Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-8, Issue-3, March 2026
DOI: 10.53469/jrse.2026.08(03).07

Transformer-based Modulation Recognition Algorithm with Multi-domain Feature Fusion

Yunpeng Pei, Guoqing Jia

Qinghai Minzu University, Xining 810007, Qinghai, China

Abstract: ***In low signal-to-noise ratio environments, the performance of modulation recognition is severely affected by noise, making it hard to boost the overall recognition rate. What's more, existing modulation signal recognition algorithms based on a single feature can't adapt to the different signal characteristics under varying signal-to-noise ratios. To solve these problems, this paper proposes a Transformer-based modulation recognition algorithm with multi-domain feature fusion. This algorithm extracts three types of signal features at the same time: I/Q time-domain features, constellation diagram features and instantaneous features. It uses the Transformer self-attention mechanism to automatically learn the contribution of different features under different signal quality levels and generate weight scores for each feature. Each score represents how important the corresponding feature is to the final classification decision. The original generated scores are normalized into a probability distribution via Softmax to get the final attention weights, which are then used to perform a weighted summation of the three feature sets—each feature set is multiplied by its corresponding attention weight. The fused final feature set is then used for classification. Experimental results show that the method proposed in this paper has obvious advantages in the low signal-to-noise ratio range of -10dB to -4dB. It outperforms the LSTM model (with a close overall recognition rate) by 6.62%, and its overall recognition rate is 12.13%, 6.81%, 1.45% and 5.19% higher than that of CNN2, ResNet, LSTM and MCformer respectively.***

Keywords: Automatic Modulation Recognition, Deep Learning, Transformer, Self-Attention Mechanism, Feature Fusion.

1. Introduction

With the rapid development of information technology, wireless communication technology has become an indispensable infrastructure in modern society. Starting from the digitalization of the 2nd Generation mobile communication system (2G), to the full commercialization of 5G networks and the gradual clarification of the 6G vision today, the field of wireless communication is undergoing an unprecedented transformation. In this evolutionary process, to meet the communication demands of high rate, low latency, massive connections and high reliability, modulation technology – a core component of communication systems – is trending toward increasing complexity and diversification. From traditional analog modulation (e.g., AM, FM) to digital modulation (e.g., BPSK, QPSK, QAM, OFDM, etc.), and further to Adaptive Modulation and Coding (AMC) technology that adapts to complex channels, the selection of modulation methods directly determines the spectral efficiency and transmission performance of communication systems.

Automatic Modulation Recognition (AMR) refers to judging the modulation mode adopted by a received signal sample through analysis of it [1-3]. It is a key link in communication signal processing, widely applied in both military and civilian fields, and boasts extremely high strategic value. In the civilian field, AMR is one of the core technologies for realizing cognitive radio. Cognitive radio requires terminals to sense the surrounding spectrum environment and intelligently adjust transmission parameters to access idle frequency bands [4]. In this process, accurate AMR can help secondary users identify primary user signals, avoid harmful interference [5], and achieve efficient reuse of spectrum resources. In addition, AMR serves as a prerequisite for judging signal legality in regulatory work such as radio spectrum management, interference source location, and illegal signal monitoring (e.g., cracking down on "fake base stations" and "illegal radio stations"). In the military field, AMR is the "eyes" of electronic warfare systems. In modern information-based warfare, seizing electromagnetic dominance is the key to gaining the initiative in wars. After intercepting enemy communication signals, it is necessary to first identify their modulation modes to further demodulate the signals for intelligence gathering, or conduct targeted jamming and deception. Therefore, the accuracy and real-time performance of AMR technology are directly related to the acuteness of battlefield situational awareness and the effectiveness of electronic countermeasures.

Traditional modulation recognition methods are mainly divided into two categories: decision theory-based methods [6] and statistical pattern recognition methods based on feature extraction. The former relies on specific signal parameters and expert knowledge; while it features strong interpretability, it has high computational complexity [7], is sensitive to the Signal-to-Noise Ratio (SNR), and struggles to handle unknown or non-standard signals [8][9]. The latter mainly depends on manually designed features by experts (such as high-order cumulants [10], cyclic spectrum features, instantaneous features [11], etc.). Though its computational load is relatively low, the quality of feature design is highly reliant on expert experience [12]. Moreover, in low SNR environments, manually designed features often fail to capture the deep intrinsic laws of signals, resulting in limited generalization ability.

In recent years, the breakthrough progress of deep learning technology has opened up new avenues for modulation recognition. Models such as Convolutional Neural Networks (CNNs) [13] and Recurrent Neural Networks (RNNs) [14] have been widely applied in the field of signal processing. However, existing deep learning-based methods still have many shortcomings. For instance, merely using the I/Q two-channel time-series information of signals [15] preserves the integrity of raw data but often overlooks the spatial distribution characteristics of signals in constellation diagrams; in addition, CNNs focus mainly on local features

[16] and struggle to capture long-range temporal dependencies. While the sole use of constellation diagrams is intuitive [17-19], it tends to lose the time-series information of phase and amplitude during image conversion. Furthermore, most existing studies focus on the extraction of single features and lack an effective fusion mechanism for multi-modal and multi-level features. Therefore, designing a deep learning model that can fully explore the multi-domain features of signals and achieve their efficient fusion is of great significance for improving the performance of modulation recognition.

2. Related Work

Under different signal conditions, the importance of signal I/Q time-domain features, constellation features, and instantaneous features varies. For example, constellation features are clearer at high SNR and should be assigned higher weights; instantaneous features are more reliable at low SNR and should be assigned higher weights. Adaptive feature fusion uses the attention mechanism to automatically learn this dynamic weight allocation strategy. It increases the proportion of more reliable feature weights under different signal conditions, thereby improving the recognition rate of the algorithm.

2.1 Overall Framework of the Algorithm

The overall flow of the multi-domain feature fusion algorithm is as follows: the I/Q time-domain features, constellation features, and instantaneous features of the signal are extracted separately using appropriate algorithms, and then fed into the adaptive fusion module for adaptive weighted fusion. The resulting fused features are sent to the Transformer encoder for positional encoding, and then input into the classifier for classification. The detailed flow is shown in Figure 1.

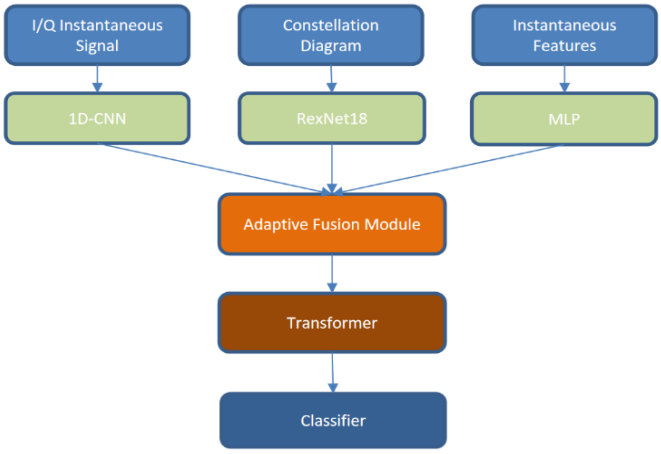

Figure 1: Flowchart of the multi-domain fusion algorithm

2.2 Three Feature Extraction Branches

The I/Q time-domain signal is a one-dimensional time series, where each time instant contains sampled values of two channels: I and Q. The modulation characteristics of the signal are reflected in the temporal patterns, such as phase transitions in BPSK, amplitude variations in QAM, frequency variations in FM, etc. Therefore, the extractor must be able to efficiently process one-dimensional time-series data. To verify which algorithm is more suitable for extracting I/Q instantaneous features, this paper selects three algorithms: 1D-CNN, LSTM, and MCformer, and conducts comparative experiments on ten modulation signals from the RML2016.10a dataset. The recognition results and parameter quantities are shown in Table 1.

Table 1: Recognition Accuracy and Number of 1D-CNN, LSTM, and MCformer

model	Accuracy	Number
1D-CNN	82.1%	0.8M
LSTM	79.6%	1.2M
MCformer	83.8%	4.2M

The core of LSTM is to process the sequence step by step over time. The computation at each step depends on the hidden state of the previous step, so parallel computation is not possible. As a result, LSTM requires 128 iterations to complete one forward pass, leading to slow model training and high inference latency. Although MCformer achieves 1.7% higher accuracy than 1D-CNN, its parameter count is more than five times that of 1D-CNN. In practical deployment, its inference latency is much higher than 1D-CNN. Compared with the 1.7% accuracy gain, the fivefold increase in parameters is not cost-effective. With only 0.8M parameters while maintaining satisfactory accuracy,1D-CNN features fast training and inference speed, making it highly suitable for real-time applications such as signal modulation recognition.

A constellation diagram is a 2D image representation converted from I/Q signals. Different modulation schemes exhibit different dot distributions (spatial structures) in the image. The extractor needs strong image feature extraction capabilitiesto recognize spatial features such as the shape, density, and distribution of the constellation points. To verify which algorithm is more suitable for extracting constellation features, this paper selects ResNet18, MobileNetV2, and a simplified ResNet50for comparative experiments on ten modulation signals from the RML2016.10a dataset. The recognition results and parameter counts are shown in Table 2.

Table 2: Recognition Accuracy and Number of ResNet18, MobileNetV2 and ResNet50

model	Accuracy	Number
ResNet18	80.4%	11M
MobileNetV2	76.9%	3.4M
ResNet50	81.0%	25M

Although ResNet50 has a much larger number of parameters, its recognition accuracy is not significantly improved, and its deployment cost is considerably higher. MobileNetV2 greatly reduces the number of parameters, but its accuracy drops by 3.5% compared with ResNet18. After balancing performance and efficiency, this paper finally selects ResNet18 as the feature extraction network for constellation diagrams.

The instantaneous features are 12-dimensional statistics manually extracted from I/Q signals, which are already highly compressed numerical features without temporal or spatial structure. They have a low input dimension (12 dimensions) and a relatively sufficient number of samples. Therefore, the extractor only needs to perform a nonlinear mapping from low dimension to high dimension, without requiring complex structural inductive bias. Thus, this paper adopts a simple, easy-to-implement and effective 3-layer MLP network.

2.3 Adaptive Feature Fusion Module

The adaptive feature fusion module receives 256-dimensional feature vectors (f_IQ, f_Const, f_Inst) output by the three feature extractors, and performs intelligent fusion through the attention mechanism. It outputs a 256-dimensional fused feature vector for the classifier. The structure of this module is shown in Figure 2.

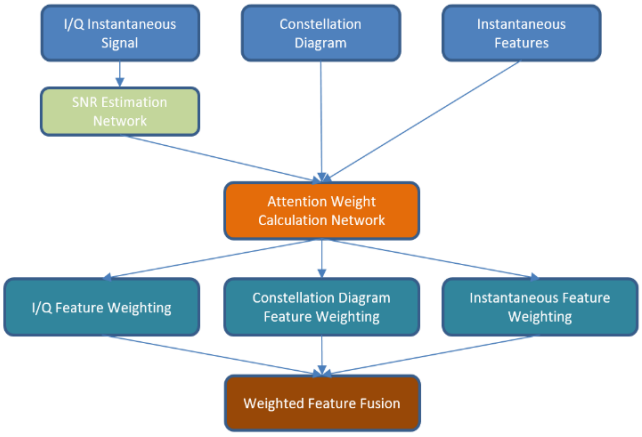

Figure 2: Structure of the Adaptive Fusion Module

The I/Q signal is the most original signal representation and contains the most complete signal quality information. Using the I/Q features as the input of the SNR estimation network can most accurately reflect the current SNR level. The SNR estimation network outputs a quality index between 0 and 1: a value close to 1 indicates high SNR (good signal quality), and a value close to 0 indicates low SNR (poor signal quality). This estimated value is used to dynamically adjust the weights of the three feature branches. Generating the weighted features can be divided into three steps.

Step 1: After the three feature branches are input into the attention weight calculation network, feature concatenation is performed. The concatenated vector contains complete information of all three features, providing a global perspective for the attention network. This allows the network to access all features and observe information from three different perspectives simultaneously.

Step 2: The concatenated features are fed into the first fully connected layer, which reduces the 768-dimensional high-dimensional features to 128 dimensions. This reduces computation, filters out redundant information, and extracts key features. Meanwhile, the weight matrix of the fully connected layer automatically learns the correlations among the three feature branches. The ReLU activation function is applied to the dimension-reduced features to introduce nonlinearity, enabling the model to learn more complex patterns and enhance representation ability.

The dimension-reduced features are fed into the second fully connected layer to generate weight scores. Through the weight matrix of the fully connected layer, the model automatically learns the contribution of different features under varying signal quality. Specifically, the 128-dimensional intermediate features are mapped to 3-dimensional raw scores. Each score represents the importance of the corresponding feature branch (I/Q time-domain feature, constellation feature, instantaneous feature) to the final classification decision. For example, under high SNR conditions, the I/Q time-domain feature and constellation feature will obtain higher scores. The specific expression is as follows:

$$s = W_2 \cdot z + b_2 \tag{1}$$

$$s = [s_{iq}, s_{const}, s_{inst}] \tag{2}$$

W_2 is the weight matrix with a dimension of 3×128, z is the intermediate feature, b_2 is the bias vector with a dimension of, $s_{iq}, s_{const}, s_{inst}$ denote the raw scores of the three feature branches, respectively.

The generated raw scores are converted into a probability distribution via Softmax normalization. The specific expression is as follows:

$$\partial_i = \frac{exp\,(s_i)}{\sum_j exp\,(s_j)} \tag{3}$$

s_i denotes the raw score of the i-th feature branch, $\exp(s_i)$ maps each score to the positive domain by taking the natural exponential, $\sum_j \exp(s_j)$ is the sum of exponentials of all scores for normalization, ensuring the sum of all weights equals 1.

Step 3: The generated attention weights are used to perform weighted summation on the three feature branches. Each feature is multiplied by its corresponding attention weight, so as to increase the weight of important features and reduce the weight of secondary features. The specific expression is as follows:

$$f_{fused} = \sum_i \partial_i \cdot f_i = \partial_{iq} \cdot f_{iq} + \partial_{const} \cdot f_{const} + \partial_{inst} \cdot f_{inst} \tag{4}$$

f_{fused} denotes the weighted fused feature, f_i denotes the i-th feature branch.

3. Experimental Design and Result Analysis

3.1 Experimental Setup and Dataset

The open benchmark dataset RML2016.10a is adopted in this paper. Since WBFM (Wideband Frequency Modulation) is a wideband frequency modulation scheme, its spectrum presents a global wideband continuous distribution feature, and there are certain silent regions in the signal. Only 128 sampling points are used for sampling in the dataset, and the signal length is short, which makes it difficult for the model to effectively capture its features. Therefore, ten modulation signals are selected, including 8PSK, DSB, SSB, BPSK, CPFSK, GFSK, PAM4, QAM16, QAM64, and QPSK. The signal-to-noise ratio (SNR) ranges from −10 dB to +18 dB with an interval of 2 dB, totaling 1.5×10^9 data samples. The dataset is split into training set, validation set, and test set with a ratio of 6:2:2 for training and testing.

The Adam optimizer is selected, with the learning rate set to 0.001, the batch size set to 512, and the training process conducted for 100 epochs. During the experiment, PyTorch is used to build the neural network for training. The computer configuration is as follows: Intel Core i5-13600KF CPU and NVIDIA GeForce RTX 4070Ti Super GPU.

3.2 Algorithm Performance Analysis

The average recognition accuracy of the proposed method on the dataset under different signal-to-noise ratios (SNR) and the recognition accuracy of each type of signal are shown in Figure 3.

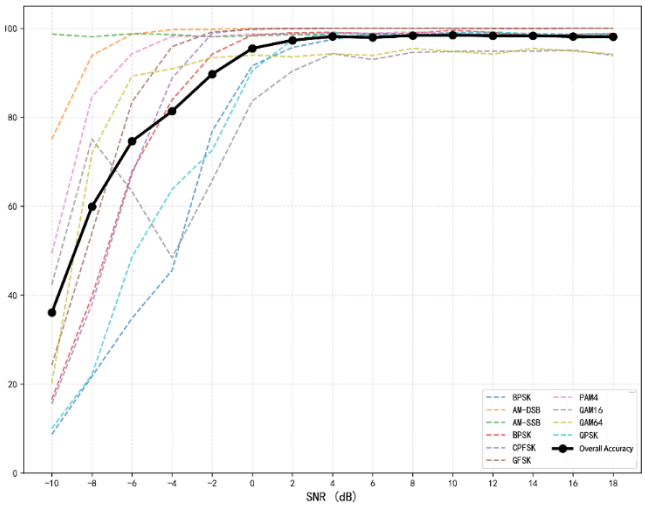

Figure 3: Recognition Curve of Each Signal Type of the Proposed Method Under Different Signal-to-Noise Ratios (SNR)

It can be seen from Figure 3 that with the increase of signal-to-noise ratio (SNR), the recognition accuracy of almost every type of signal increases and fluctuates slightly under high SNR. When SNR ≥ 2 dB, the overall recognition accuracy of the algorithm can be maintained above 97%; when SNR = 0 dB, the overall recognition accuracy is 95.5%; and the average recognition accuracy in the low SNR interval of −10 dB ≤ SNR ≤ 0 dB is 68.3%. Four types of signals, namely AM-DSB, AM-SSB, GFSK, and PAM4, can still maintain high recognition accuracy under low SNR. Among them, the accuracy of AM-SSB remains above 98% under the full SNR range; the accuracy of AM-DSB is still 94% when SNR = −8 dB; the accuracy of GFSK is 84% when SNR = −6 dB; and the accuracy of PAM4 can still maintain 85% when SNR = −8 dB. Signals including 8PSK, QPSK, BPSK, QAM16, QAM64, and CPFSK (corrected from CFPSK) are sensitive to changes in SNR. In the low SNR interval of SNR ≤ −4 dB, their accuracy decreases rapidly with the decrease of SNR.

3.3 Comparison with Baseline Models

In this paper, the proposed model is compared in performance with baseline models including CNN2, ResNet, LSTM, and MCformer, with the batch size uniformly set to 512. Table 3 shows the overall classification and recognition accuracy of the baseline models and the proposed model within the SNR range of −10 dB to +18 dB, and the corresponding relationship between the recognition accuracy of different networks and SNR is also presented in Table 3.

To more intuitively compare the recognition accuracy of the proposed method and various baseline models under different signal-to-noise ratios (SNR), the overall recognition rate curve in Figure 4 is plotted.

Table 3: Recognition Accuracy Table

SNR (dB)	CNN2 (%)	LSTM (%)	MCformer (%)	ResNet (%)	Ours (%)
-10	15.55	24.21	24.63	40.23	36.10
-8	33.20	48.29	46.58	55.34	59.91
-6	61.60	70.81	65.29	69.32	74.63
-4	74.30	82.60	75.36	76.87	81.38
-2	81.61	91.76	85.74	83.01	89.71
0	84.87	96.99	91.53	86.35	95.50
2	86.81	97.83	93.60	87.50	97.33
4	87.72	98.22	94.69	89.04	98.17
6	87.32	98.28	94.82	89.72	97.97
8	87.95	98.23	95.14	89.95	98.37
10	88.25	98.37	94.97	90.08	98.45
12	87.44	98.36	95.03	90.34	98.45
14	87.16	98.36	95.06	89.80	98.32
16	86.86	98.29	95.03	90.10	98.18
18	87.84	98.16	95.09	90.65	98.10

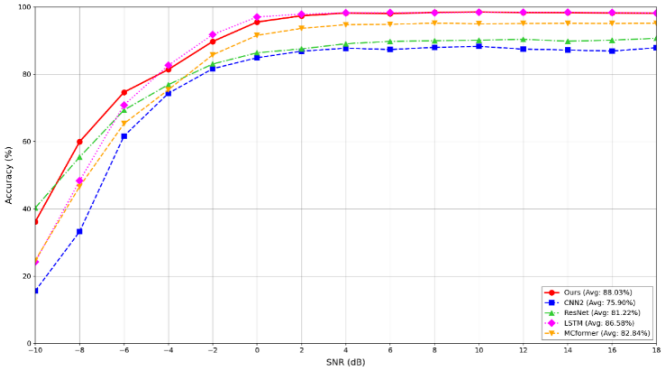

Figure 4: Overall Classification Recognition Curve of the Five Methods

It can be seen from Figure 4 that the average recognition accuracy of the proposed model in this paper reaches 63.01% in the interval of −10 dB to −4 dB, which is 6.62% higher than the average recognition accuracy of LSTM (56.48%). After −4 dB, with the increase of SNR, the recognition capabilities of the two models are equivalent. ResNet is only slightly higher than the proposed model at −10 dB, and is much lower than the proposed model at other SNR values. The recognition rates of CNN2 and MCformer are far lower than that of the proposed model under all SNR conditions. In terms of overall recognition rate, the proposed model reaches 88.03%, which is 12.13% higher than the CNN2 model, 6.81% higher than the ResNet model, 1.45% higher than the LSTM model, and 5.19% higher than the MCformer model.

4. Conclusion

Aiming at the problem that models such as CNN2, ResNet, LSTM, and MCformer use a single feature for recognition, which cannot adapt to the characteristics of different signals under different signal-to-noise ratios, this paper extracts three types of features—IQ time-domain features, constellation diagram features, and instantaneous features—and performs adaptive feature fusion. By generating different attention weight parameters, dual dynamic feature selection of SNR adaptability and modulation type adaptability is realized to improve recognition accuracy. Experiments show that the proposed method has obvious advantages between −10 dB and −4 dB, which is 6.62% higher than the LSTM model with a similar overall recognition rate. In terms of overall recognition rate, it is 12.13%, 6.81%, 1.45%, and 5.19%

higher than CNN2, ResNet, LSTM, and MCformer, respectively, thus verifying that the proposed method has good classification performance.

References

[1] Zhang X, Lu G, Wang J, et al. Efficient Modulation Recognition with Minimal Samples Leveraging Architecture Search and Knowledge Transfer in Combined Radar-Communication Environments [C]// 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring). IEEE, 2024: 1-5.

[2] Zhang W, Xue K, Yao A, et al. Automatic modulation recognition based on multimodal information processing: A new approach and application[J]. Electronics, 2024, 13(22): 4568.

[3] Kim K. A deep learning method for automatic modulation recognition in the time--frequency domain [J]. 2025.

[4] Vithalani A, Shah C. Lightweight Multi-Channel Gated Recurrent Deep Neural Network for Automatic Modulation Recognition in Spatial Cognitive Radio[J]. Applications of Modelling and Simulation, 2024, 8: 26-39.

[5] El-haryqy N, Madini Z, Zouine Y. Radio frequency interference detection and automatic modulation recognition based on mask RCNN[J]. Int. J. Comput. Netw. Commun, 2024, 16(5): 23-42.

[6] KIM K, POLYDOROS A. Digital modulation classification: the BPSK versus QPSK case [C]. MILCOM 88, 21st Century Military Communications - What's Possible?'. Conference record. Military Communications Conference, San Diego, CA, USA, 1988: 431-436.

[7] SAPIANO P C, MARTIN J D. Maximum likelihood PSK classification using the DFT of phasehistogram[C]. Proceedings of GLOBECOM'95, Singapore, 1995: 1029-1033.

[8] SILLS J A. Maximum-likelihood modulation classification for PSK/QAM[C]. MILCOM 1999.IEEE Military Communications. Conference Proceedings, Atlantic City, NJ, USA, 1999: 217-220.

[9] MARCHAND P, LE MARTRET C, LACOUME J L. Modulation classification based on amaximum - likelihood receiver in the cyclic-HOS domain[C]. 9th European Signal ProcessingConference, Rhodes, Greece, 1998: 1-4.

[10] GHANI N, LAMONTAGNE R. Neural networks applied to the classification of spectralfeatures for automatic modulation recognition[C]. Proceedings of MILCOM'93-IEEE MilitaryCommunications Conference, Boston, MA, USA, 1993: 111-115.

[11] NANDI A K, AZZOUZ E E. Automatic analogue modulation recognition[J]. Signal processing,1995, 46(2): 211-222.

[12] AZZOUZ E E, NANDI A K. Automatic identification of digital modulation types[J]. Signalprocessing, 1995, 47(1): 55-69.

[13] HUYNH-THE T, HUA C H, PHAM Q V, et al. MCNet: An efficient CNN architecture for robustautomatic modulation classification[J]. IEEE Communications Letters, 2020, 24(4): 811-815.

[14] KE Z, VIKALO H. Real-time radio technology and modulation classification via an LSTMauto-encoder[J]. IEEE Transactions on Wireless Communications, 2022, 21(1): 370-382.

[15] O'Shea T J, Corgan J, Clancy T C. Convolutional radio modulation recognitionnetworks [C]// Engineering Applications of Neural Networks: 17th International Conference, EANN2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17. Springer International Publishing, 2016:213-226.

[16] Chen Z, Cui H, Xiang J, et al. SigNet: A novel deep learning framework for radio signalclassification[J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 8(2):529-541.

[17] Wang D, Zhang M, Li Z, et al. Modulation format recognition and OSNR estimation usingCNN-based deep learning[J]. IEEE Photonics Technology Letters, 2017, 29(19): 1667-167

[18] Mao Y, Zhu M L, Sun T, et al. Automatic modulation classification based on snr estimation viatwo-stage convolutional neural networks[C] //2021 6th International Conference on IntelligentComputing and Signal Processing (ICSP). IEEE, 2021: 294-29

[19] Peng S, Jiang H, Wang H, et al. Modulation classification based on signal constellation diagramsand deep learning[J]. IEEE transactions on neural networks and learning systems, 2018, 30(3):718-727.