AI Engineering
https://www.sciltp.com/journals/aieng

Article

Multimodal Deep Learning based Automatic Modulation Recognition: Fusion of Signal Modalities

Xin Lin [1], Qinggeng Guo [1], Xi Yang [2], Ping Wu [3] and Shengliang Peng [1,*]

[1] College of Information Science and Technology, Huaqiao University, Xiamen 361021 , China
[2] College of Communication and Electronic Engineering, Jishou University, Jishou 416000, China
[3] Department of Electrical Engineering, Uppsala University, SE-751 05 Uppsala, Sweden
* Correspondence: peng.shengliang@hqu.edu.cn

How To Cite: Lin, X.; Guo, Q.; Yang, X.; et al. Multimodal Deep Learning based Automatic Modulation Recognition: Fusion of Signal Modalities. *AI Engineering* **2026**, *2*(1), 5. https://doi.org/10.53941/aieng.2026.100005

Received: 6 March 2026
Revised: 19 May 2026
Accepted: 4 June 2026
Published: 23 June 2026

Abstract: Automatic modulation recognition (AMR) has been becoming an indispensable part in intelligent communications systems, especially for cognitive radio and radio regulation. With the fast development of machine learning in the recent years, deep learning (DL) has been applied to AMR. However, existing DL based AMR methods only rely on a single signal modality, which limits the full utilization of signal features and restricts recognition performance. Thus, this paper proposes to develop multimodal DL based AMR, which is established, in particular, by exploring a variety of modalities to represent the received signals (e.g., in-phase and quadrature sequences, constellation diagram), and then fusing two or more of the modalities at three different stages of the DL architecture. The three stages of the multimodal fusion include the early fusion, intermediate fusion and late fusion in which the multiple signal modalities are fused before, in between and after the DL model, respectively. The algorithms for the three fusion methods are proposed and implemented in experiments. Evaluation of the algorithms is made according to accuracy, complexity and flexibility. The results show that (1) the early fusion exhibits satisfactory classification accuracy, least complexity and not good flexibility; (2) the intermediate fusion gives best accuracy, high complexity and satisfactory classification flexibility; and (3) the late fusion gives low accuracy, high complexity and best flexibility. Moreover, the proposed multimodal DL based AMR algorithms consistently outperform single signal modality approaches under tested conditions, including different DL model structures, sample quantities, and channel models, demonstrating strong generality and universal superiority for automatic modulation recognition.

Keywords: automatic modulation recognition; deep learning; multimodal; fusion

1. Introduction

Signal modulation is a fundamental process in wireless communications, converting digital or analog data into radio waves. With the emergence of cognitive radio (CR) and intelligent communications systems, automatic modulation recognition (AMR) has become attractive to researchers. For instance, in non-cooperative communications, AMR is utilized for spectrum interference monitoring, radio fault detection, dynamic spectrum access, and opportunistic mesh networking, achieving the integration of communications and sensing [1,2]. In cognitive radio, it is used to aid spectrum sensing, enhancing the efficiency of spectrum utilization [3]. In radio regulation, it is beneficial for the timely monitor and discovery of illegal transmitters, which facilitates equitable access to and rational use of the radio resource [4]. Moreover, in the field of military communications, AMR is employed to analyze the specific parameters of the captured signals during the electromagnetic spectrum warfare [5].

With the quick development of machine learning, deep learning (DL) has gained great attention from both the academia and industry. Consequently, in recent years, AMR based on DL has been becoming significantly prosperous.

In [6], an uncertainty-aware AMC framework based on supervised contrastive learning is proposed to effectively enhance the recognition accuracy and generalization robustness of modulation recognition under complex open-world scenarios. In [7], the first domain-incremental learning (DIL) paradigm for SMC is proposed, and a parameter-efficient isolation DIL method is designed to achieve stable incremental learning for modulation classification. In [8], convolutional neural networks (CNNs) are used to extract features from in-phase and quadrature (IQ) signals to achieve AMR. In [9], a lightweight AMR network is proposed by combining two types of sparse convolutional layers, namely depthwise and regular layers, within a single architecture to achieve a high recognition accuracy. A capsule-based network with higher classification accuracy and fewer signal samples is introduced in [10]. The deep sparse-filtering CNNs are explored for AMR to improve the recognition accuracy and robustness [11]. In [12], the long short-term memory (LSTM) network, a special type of recurrent neural network, is suggested to complete AMR. In [13], CNNs and gated recurrent units are employed to achieve higher accuracy while significantly reducing the number of model parameters in AMR. In [14], a modified AMR method termed the frame-wise embedding aided transformer is proposed to capture global features of the signal.

Multimodal DL is one of the main branches of DL. It is used to fuse multimodal representations of data to deal with complex cases, and has found many applications in various fields, such as bioinformation, image processing and automatic control [15]. For example, in [16], deep Boltzmann machine that intermediately fuses three modalities of gene expression, deoxyribonucleic acid methylation, and drug response is applied to address the cancer subtype clustering problem. In [17], a multispectral imagery modality is employed with an early fusion method, as well as a fully connected neural network, to solve the semantic segmentation problem. Both image and optical flow modalities are combined using a late fusion approach based on CNNs for action recognition as proposed in [18]. The authors of [19] treats video features, GPS coordinates, and vehicle dynamics as three modalities that are processed via an LSTM and an intermediate fusion method to tackle the driver activity anticipation problem.

As for AMR, a modulated signal to be recognized can be preprocessed into various formats which are regarded as different signal modalities [20]. However, in the existing works, single signal modality is mostly used for AMR. In [8], the modulated signals are represented as the modality of IQ sequence to distinguish 3 modulation types of binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), and eight phase shift keying (8PSK). The authors of [21] suggested utilizing the signal modality of amplitude histogram sequences to differentiate 3 modulation types of QPSK, 16 quadrature amplitude modulation (16QAM) and 64 quadrature amplitude modulation (64QAM). In [22], the fast Fourier transform of signal sequence is applied as input, and a four-layer CNN is used for convolution and pooling operations. Constellation diagrams are also widely-used signal modalities for automatic feature selection to achieve higher accuracy in AMR [23]. In [24], the contour stellar image is used for AMR recognition to address challenges associated with waveforms in the physical layer. In [25], the modulated signal is first converted into the eye diagram, and then recognized by CNNs. Most recently, some studies begin to employ multiple signal modalities to obtain higher AMR accuracy. In [26], signals are represented using both signal modalities of IQ sequence and constellation diagram, and a hierarchical recognition architecture with CNNs is employed to achieve AMR. In [27], waveform and spectrum signal modalities are fed into a deep residual network (ResNet) for AMR. The features of modulated signal are extracted from both the image as well as IQ waveform and tackled with kernel principal component analysis [28]. In [29], joint utilization of signal modalities in the time and frequency domains is discussed, and a domain adversarial network is employed for the few-shot semi-supervised AMR. In [30], one-dimensional signal features and constellation diagram features are combined through wavelet transforms to form a feature set, which is then filtered using reinforcement learning to obtain a small subset of features for AMR. It should be pointed out that, although multiple signal modalities are involved in these works, multimodal DL has not been fully applied to aid AMR, especially considering that multiple signal modalities can be fused either in the early, intermediate or late phase.

This paper is aimed for a comprehensive study on multimodal DL based AMR, including the methods for fusing multiple signal modalities. The main contributions of this paper are summarized as follows:

- We propose a multimodal DL based algorithm for AMR, in which the features of modulated signal are extracted from two signal modalities of IQ sequence and constellation diagram and learned by either the SigCNN or ResCNN model to achieve objective evaluation of different fusion methods.
- We investigate three distinct fusion methods, including early fusion, intermediate fusion and late fusion to effectively utilize multiple signal modalities of the modulated signal.
- We evaluate the performance of the proposed algorithm with different fusion methods under various scenarios of network architectures, sample numbers and channel conditions.

The remaining of this paper is organized as follows. The problem of AMR based on DL is formulated in Section 2. Then in Section 3 the approaches to representing the modulated signal into different modalities are

introduced. In Section 4 the algorithms for the multimodal DL and the methods for fusing the signal modalities are discussed in detail. In Section 5 the experiment and the results for evaluating the algorithm in various scenarios are presented and discussed. Finally in Section 6 the conclusions are made.

2. Problem Formulation

A typical DL based AMR system is illustrated in Figure 1. It is composed of three parts: signal receiving, feature extraction and modulation recognition. In the first part, the modulated signal corrupted by the channel impact and additive noise is received for AMR. In the second part, the received signal is represented into a proper signal modality, whose features are extracted by the convolutional or other hidden layers. In the last part, a classifier is constructed to handle the extracted features and deduce the modulation type of the received signal.

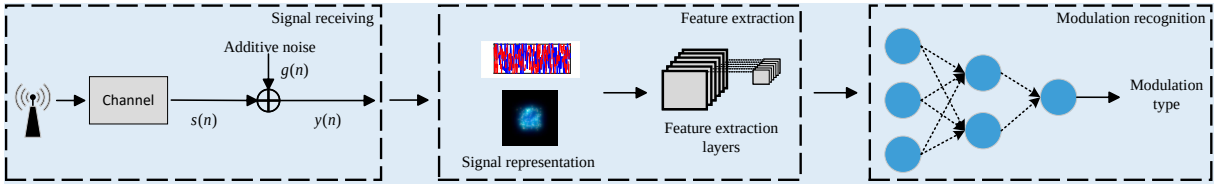


Figure 1. System of AMR based on DL.

Specifically, the received signal in this system is expressed as

$$y\left(n\right)=s\left(n\right)+g\left(n\right), \tag{1}$$

where $g(n)$ is the additive white Gaussian noise (AWGN), and $s(n)$ is the modulated signal with channel impacts,

$$\begin{aligned} s(n) = & Ae^{j(2\pi f_0 nT+\theta_n)} \\ & \cdot \sum_{l=-\infty}^{\infty} x(l)h(nT-lT+\epsilon_T T), \end{aligned} \tag{2}$$

where $x\left(l\right)$, A and f_0 are the symbol sequence, amplitude factor and residual frequency offset, respectively. θ_n represents the phase jitter, T represents the symbol spacing, $h\left(\cdot\right)$ represents the residual channel effects, and ϵ_T $(0 \leq \epsilon_T \leq 1)$ represents timing errors.

In DL based AMR, deep neural networks (DNNs) are utilized to extract signal features and classify modulation types. Given a batch of samples of the received signal, our task is to recognize the modulation type C_m $(1 \leq m \leq M)$ adopted by the modulated signal $x(l)$ from a candidate set of modulation types $\{C_1, C_2, \ldots, C_M\}$. For simplicity, this research considers five prevalent and representative modulation scheme as examples to illustrate the idea of multimodal fusion, including three low-order schemes of BPSK, QPSK as well as 8PSK and two high-order QAM with the orders of 4 and 6.

3. Multimodal Signal Representation

Signal representation is a critical technique that converts the signal into proper formats before feeding it into DNNs. According to [20], various representations have been suggested and each representation can be regarded as a signal modality. Similar to [31], two popular signal modalities, the IQ sequence and the constellation diagram, which are used to capture different aspects of the signal, are considered as examples. The former captures 1-D temporal signal variations, while the latter reflects 2-D spatial symbol distribution. The signals involved are assumed to operate in a coherent and synchronous scenario with single-tone signaling [23]. It is also assumed that carrier, timing, and waveform recovery have been accomplished.

3.1. IQ Sequence

The IQ sequence has abundant modulation information and high adaptability to different signal types, and is therefore widely used in AMR [32]. The received signal $y(n)$ can be rewritten as

$$y(n)=y_i(n)+j\cdot y_q(n), n=0,1,\ldots,N_s-1, \tag{3}$$

where $y_i(n)=Re[y(n)]$, $y_q(n)=Im[y(n)]$ represents the real and imaginary parts of the signal, respectively. Assuming N_s samples of the received signal are collected within an observation, the signal modality of IQ sequence

is given by

$$\mathbf{Y_{iq}} = \begin{bmatrix} y_i(0) & y_i(1) & y_i(2) & \dots & y_i(N_s-1) \\ y_q(0) & y_q(1) & y_q(2) & \dots & y_q(N_s-1) \end{bmatrix}, \tag{4}$$

where $\mathbf{Y_{iq}} \in \mathbb{R}^{2\times N_s}$. The visualization of IQ sequences for different modulation types is shown in Figure 2.

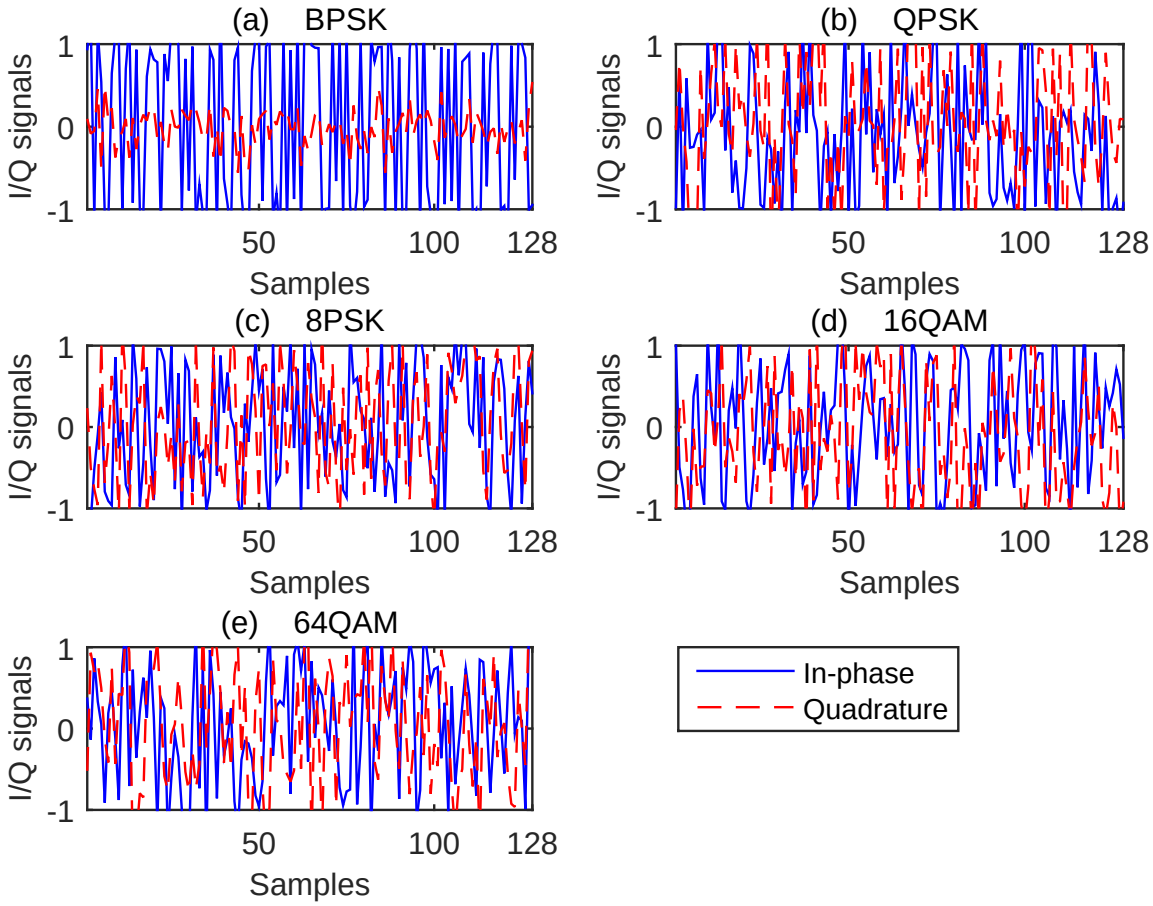


Figure 2. In-phase and quadrature signals for different modulation types: (**a**) BPSK; (**b**) QPSK; (**c**) 8PSK; (**d**) 16QAM; (**e**) 64QAM.

3.2. Constellation Diagram

Constellation diagram is a widely used 2-D image that maps signal samples to scattered points on a complex plane. According to [23], three-channel constellation diagram is explored as one of the signal modalities in this paper.

To achieve this goal, the received signal is firstly converted into enhanced gray constellation diagram, which considers the positional differences of the signal samples within a pixel and the impacts of each sample on its adjacent pixels using an exponential decay model

$$B_{u,v} = P \cdot e^{-\lambda_i \cdot d_{u,v}}, \tag{5}$$

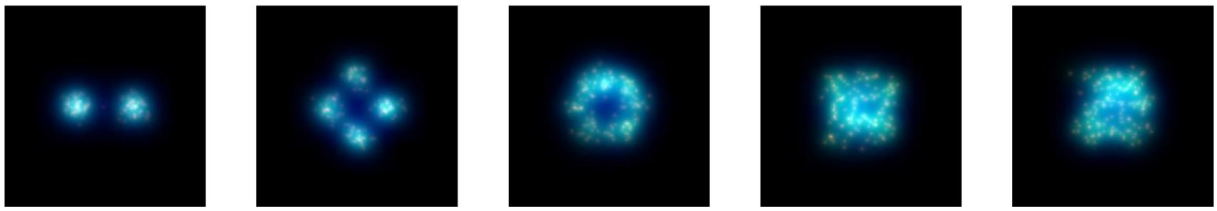
where $B_{u,v}$ represents the influence of sample point $\#u$ on Pixel $\#v$, P is the power of each sample point, $d_{u,v}$ is the centroid distance between sample point $\#u$ and Pixel $\#v$, and λ_i is the exponential decay rate. Considering that the image resolution is $\alpha \times \alpha$, the enhanced gray constellation diagram is given by

$$\mathbf{Y_{gray}} \in \mathbb{R}^{\alpha\times\alpha\times 1}. \tag{6}$$

Three enhanced gray constellation diagrams can be obtained by taking different exponential decay rates λ_1, λ_2 and λ_3 for the same batch of signal samples. The three channels merely represent different signal attenuation characteristics and do not carry RGB color-related physical meaning. Then the three-channel constellation diagram is obtained by combining three enhanced gray constellation diagrams as follows,

$$\mathbf{Y_{img}} = [\mathbf{Y_{gray1}}; \mathbf{Y_{gray2}}; \mathbf{Y_{gray3}}], \mathbf{Y_{img}} \in \mathbb{R}^{\alpha\times\alpha\times 3}. \tag{7}$$

In this paper, a 3.5×3.5 complex plane is chosen to generate the constellation diagrams, with $\lambda_1 = 1$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.1$. Some examples of the three-channel constellation diagrams for different modulation types are shown in Figure 3.



(a) BPSK (b) QPSK (c) 8PSK (d) 16QAM (e) 64QAM

Figure 3. Three-channel constellation diagrams for different modulation types at the SNR of 6dB: (**a**) BPSK; (**b**) QPSK; (**c**) 8PSK; (**d**) 16QAM; (**e**) 64QAM.

4. Multimodal DL Based AMR Algorithms

Multimodal DL based AMR is proposed to fuse multiple signal modalities and enhance recognition performance. Fusion can be conducted either before, within or after DNN. Consequently, three algorithms of multimodal DL based AMR are investigated with early fusion, intermediate fusion, and late fusion, respectively, as shown in Figure 4. In the first algorithm with early fusion, available signal modalities are firstly fused and then fed into DNN. In the second algorithm with intermediate fusion, a fusion layer is added within DNN, which fuses the signal features extracted by the lower layers of DNN. In the last algorithm with late fusion, each signal modality is handled by one DNN, and the outputs of multiple DNNs are fused to obtain the recognition result.

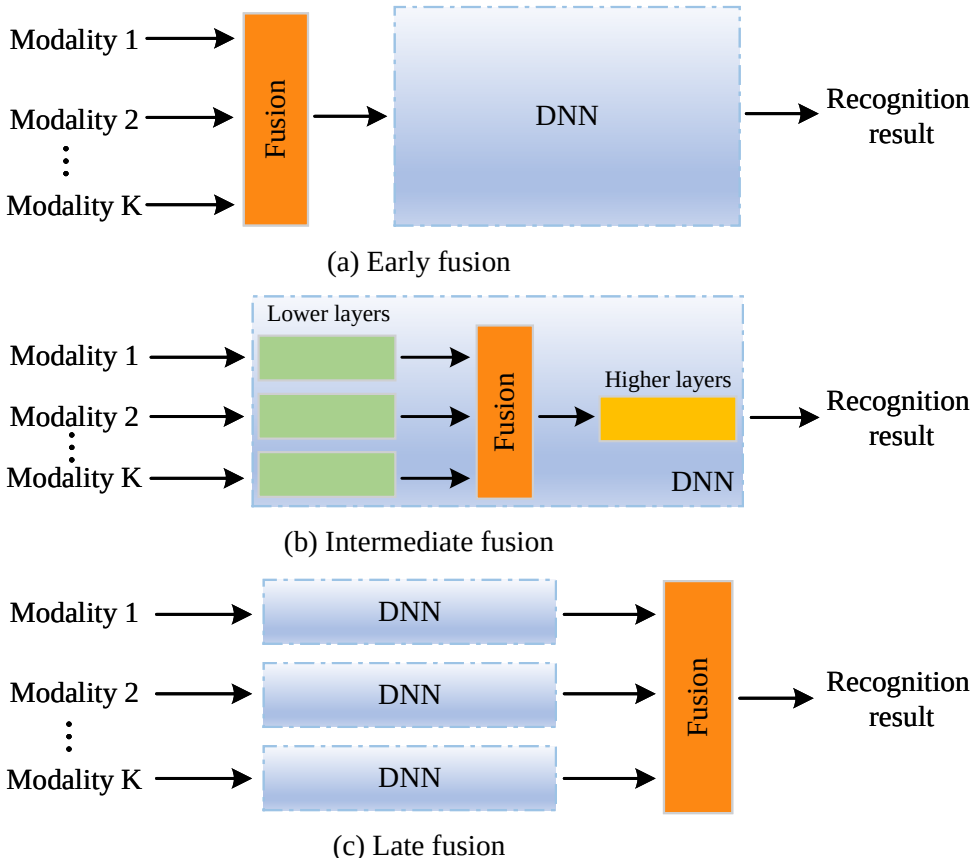


Figure 4. Different fusion stages of multimodal DL based AMR: (**a**) Early fusion; (**b**) Intermediate fusion; (**c**) Late fusion.

4.1. Algorithm with Early Fusion

The detailed steps of the proposed multimodal DL based AMR algorithm with early fusion is demonstrated in Figure 5. According to this algorithm, the received signal to be recognized is firstly converted into multiple modalities of IQ sequence and constellation diagram, which has been illustrated in Section 3. Then these signal

modalities are fused based on the early fusion method. After that, the fusion result is input into the DNN model for feature extraction and decision making.

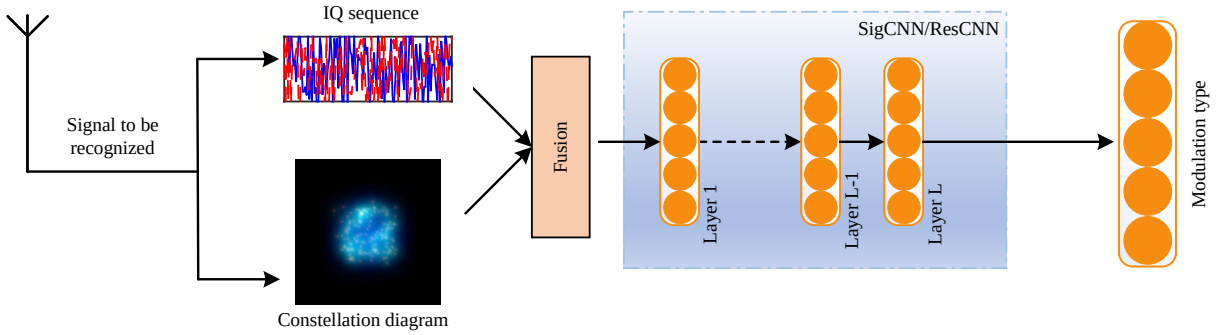


Figure 5. Proposed algorithm with early fusion.

4.1.1. Early Fusion

Early fusion is a method of integrating multiple data sources into a single data vector before being used as input in DL algorithms. This method is usually used to deal with data with different sources, types, and properties. Since the significant differences between different data sources, merging them into a single data vector without feature extraction is a challenging task. To address this issue, higher-level representations can be extracted from each modality, and the features from different data sources can be fused at only one level before inputting them into the DL algorithm. This ensures comparability of features from different data sources and ensures that the extracted features adequately reflect the information from multiple data sources.

In this paper, two signal modalities of IQ sequence and constellation diagram are considered to fuse at the early stage. Note that these two signal modalities have different data size,

$$\mathbf{Y_{iq}} \in \mathbb{R}^{2\times N_s}, \mathbf{Y_{img}} \in \mathbb{R}^{\alpha\times\alpha\times 3}, \tag{8}$$

and thus can not been fused directly. In order to solve this problem, we firstly reshape $\mathbf{Y_{iq}}$ into a vector

$$\mathbf{y_{iq}^{'}} = [y_i(0), y_q(0), ..., y_i(N_s-1), y_q(N_s-1)], \tag{9}$$

$\mathbf{y_{iq}^{'}} \in \mathbb{R}^{1\times 2N_s}$. Meanwhile, in order to facilitate multimodal fusion, we introduce a parameter denoted as α for processing the IQ sequence dimension into a new IQ signal modality matrix that aligns with the constellation diagram in both height and width, thereby enabling the two input modalities to fit the same network model and fusion module. The parameter α is determined according to N_s as follows,

$$\alpha = \left\lfloor \sqrt{2\times N_s} \right\rfloor. \tag{10}$$

Then, a new IQ signal modality matrix is constructed

$$\mathbf{Y_{iq}^{'}} = \begin{bmatrix} \mathbf{y_{iq}^{'}}(\mathbf{0}) & \cdots & \mathbf{y_{iq}^{'}}(\alpha-\mathbf{1}) \\ \vdots & \ddots & \vdots \\ \mathbf{y_{iq}^{'}}(\alpha^{\mathbf{2}}-\alpha) & \cdots & \mathbf{y_{iq}^{'}}(\alpha^{\mathbf{2}}-\mathbf{1}) \end{bmatrix}, \mathbf{Y_{iq}^{'}} \in \mathbb{R}^{\alpha\times\alpha\times 1}. \tag{11}$$

Finally, these two signal modalities are fused as follows,

$$\mathbf{X^{early}} = \left[\mathbf{Y_{iq}^{'}}; \mathbf{Y_{img}}\right], \mathbf{X^{early}} \in \mathbb{R}^{\alpha\times\alpha\times 4}. \tag{12}$$

The fused signal representation can be fed into either SigCNN or ResCNN model to achieve AMR.

It should be pointed out that the early fusion method illustrated in Equations (10) to (12) is only valid for the signal modalities of IQ sequence and constellation diagram discussed in this paper. If other signal modalities are applied, the fusion method needs to be redesigned. Moreover, if a new network model is adopted, the fusion method also needs modification to adapt to the network changes. In other words, early fusion is related to both signal modalities and network model, and lacks of flexibility. Additionally, early fusion may introduce arbitrary reshaping effects, which could be further discussed in the future work.

4.1.2. DNN Model

CNN is a type of DNNs that is widely used in computer vision, natural language processing, and other fields [33]. Compared to traditional neural networks, CNNs have the various distinct characteristics and advantages. Firstly, it uses convolution and pooling to automatically extract local features of multidimensional data, which improves computational efficiency of feature extraction. Secondly, it is superior in extracting data features, improving the recognition accuracy and robustness. Additionally, it has high generalization by stacking multiple convolutional and pooling layers to improve the performance and generalization ability of the model. This paper considers two CNN models, namely SigCNN and ResCNN, to implement multimodal DL based AMR.

SigCNN model: Similar to [34], a SigCNN model is proposed to implement multimodal DL based AMR, whose structure is detailedly shown in Figure 6. The initial layer is built with convolutional operations as well as normalization and ReLU activation. The convolutional operations employ 32 filters with the size 3×3 and utilize padding to maintain consistent output dimensions and preserve intricate details. The second layer mirrors the structure of the first, but adopts different convolutional operations with 16 filters for feature extraction. The third layer takes the operation of flattening, converting the multi-dimensional feature maps generated by the second layer into a one-dimensional vector. The fourth layer is Dropout with the rate of 0.6 to mitigate overfitting. The fifth layer takes the form of a fully-connected layer containing 128 neurons as well as ReLU activation, serving the object of transforming the previously extracted features into more abstract representations. The sixth layer once again utilizes a Dropout layer with the rate of 0.6. Finally, the last layer outputs the probability distribution vector of 5 modulation types, which can be used to determine the modulation type by choosing the type with the maximum probability.

ResCNN model: The ResCNN model is upgraded from the SigCNN by introducing two residual modules. The core idea of residual modules is to add the input of the current layer directly to the output of the subsequent layer through skip connections, which ensures the lossless transmission of information in the model, and avoids the problems of gradient disappearance and explosion [35]. Furthermore, residual modules can effectively reduce the number of network parameters, which makes it possible to construct deeper networks. Figure 7a depicts the structure of residual modules we use in ResCNN, which exploits two layers of convolutional layers with a filter size of 3×3 to extract signal features, and merges the output of the second convolutional layer with previous input through skip connections. After adding residual modules, the final structure of the ResCNN is shown in Figure 7b. There are two residual blocks: the first block uses 32 filters, and the second block uses 16 filters. The residual blocks are followed by a flatten layer, dropout (0.6), a fully-connected layer with 128 units, another dropout (0.6), and the final fully-connected output layer with 5 units. The structure of the ResCNN has two distinct differences from the SigCNN. First, the characteristic of ResCNN is adding residual module behind each convolutional layer, which allows the network to capture signal residual features. Second, the padding operation is removed in conventional operations to reduce the number of model parameters, as residual modules are powerful in extracting signal features.

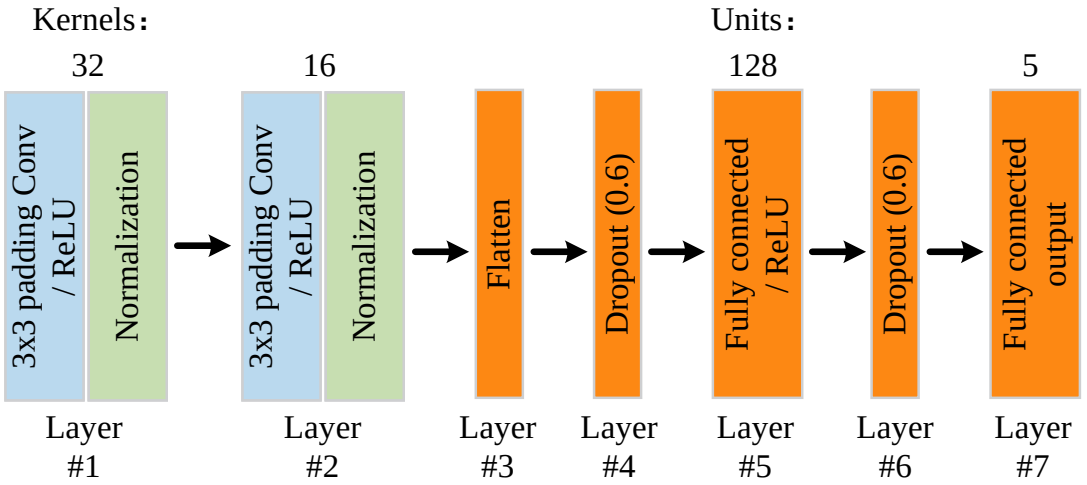
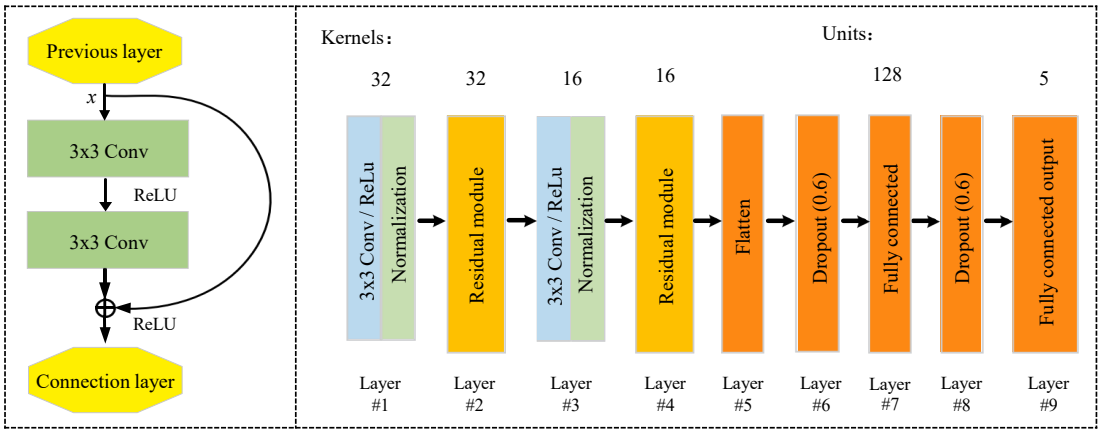


Figure 6. Network structure of SigCNN model.



(a) Residual module (b) ResCNN model

Figure 7. Residual module and ResCNN model.

4.2. Algorithm with Intermediate Fusion

The multimodal DL based AMR algorithm with intermediate fusion is illustrated in Figure 8. The received signal to be recognized is first preprocessed into multiple signal modalities of IQ sequence and constellation diagram. Similar to the proposed algorithm with early fusion, either SigCNN or ResCNN is utilized as the DNN model. However, the DNN model here comprises multiple branches of lower layers, each of which is used to extract features of one signal modality. Moreover, a fusion layer is added between the lower layers and higher layers of the DNN model to achieve intermediate fusion.

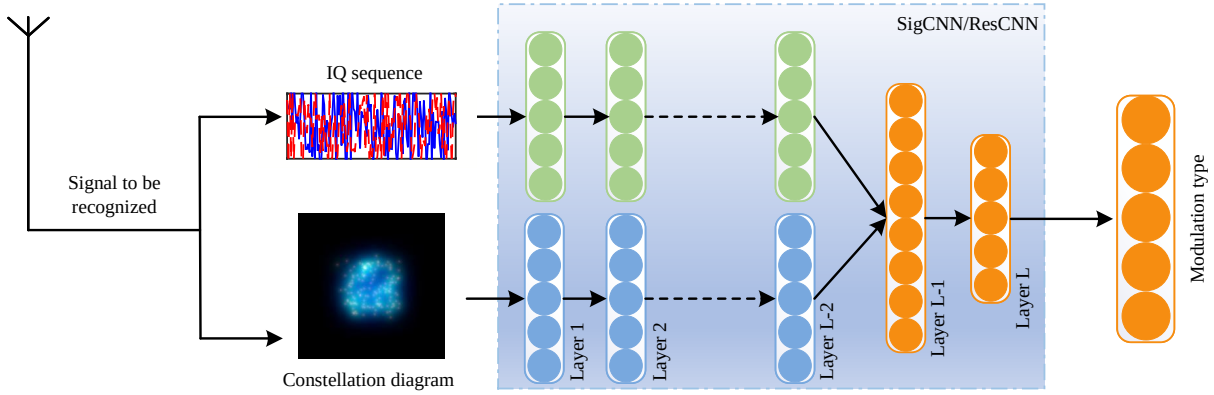


Figure 8. Proposed algorithm with intermediate fusion.

Intermediate fusion is used to generate joint features by using different features extracted from various modalities and feed them to the backend network for learning. Specifically, in multimodal DL with intermediate fusion, raw data input is firstly transform into high-level representations via lower layers of the network. Then, a fusion layer, also known as the shared representation layer, is employed to merge units of multiple modality-specific paths connected to it. After that, the merged joint representations are fed into higher layers of the network for further processing. Obviously, how to construct the fusion layer is of vital importance to intermediate fusion.

In our algorithm, the fusion layer is added preceding the final layer. For SigCNN, it is located between layers #6 and #7 in Figure 6. For ResCNN, it is inserted between layers #8 and #9 in Figure 7. The feature vectors of IQ sequences and constellation diagrams are respectively extracted by the network model

$$\mathbf{f_{iq}} = G_L^{iq}(G_{L-1}^{iq}(..G_1^{iq}(\mathbf{Y_{iq}}))), \tag{13}$$

$$\mathbf{f_{img}} = G_L^{img}(G_{L-1}^{img}(..G_1^{img}(\mathbf{Y_{img}}))), \tag{14}$$

where $G_L^{iq}(\cdot)$ and $G_L^{img}(\cdot)$ denote the transfer functions of the L-th layer for the IQ sequence and the constellation diagram network branches, respectively. Once the feature vectors $\mathbf{f_{iq}}$ and $\mathbf{f_{img}}$ are extracted from the two signal modality branches, they are fused by concatenation along the last dimension to obtain the fused feature $\mathbf{f^{inter}}$. This can be represented as

$$\mathbf{f^{inter}} = [\mathbf{f_{iq}}, \mathbf{f_{img}}], \tag{15}$$

which can be fed to the higher layer for future processing.

As shown in Equations (13) to (15), the intermediate fusion method fuses the feature vectors of multiple signal modalities instead of using the signal modalities directly. Therefore, it works well when other signal modalities are exploited. Moreover, since intermediate fusion is conducted by adding a fusion layer within the network model, the fusion method should be updated if the network model is changed. In a word, intermediate fusion is only related to the network model, and thereby has higher flexibility than early fusion.

4.3. Algorithm with Late Fusion

The multimodal DL based AMR algorithm with late fusion is illustrated in Figure 9. The signal to be recognized in this case is first converted into multiple signal modalities, including IQ sequence and constellation diagram. Then each signal modality is separately processed by one DNN model, which can be either SigCNN or ResCNN, producing an individual decision. After that, multiple individual decisions are fused according to the late fusion method to make the final decision.

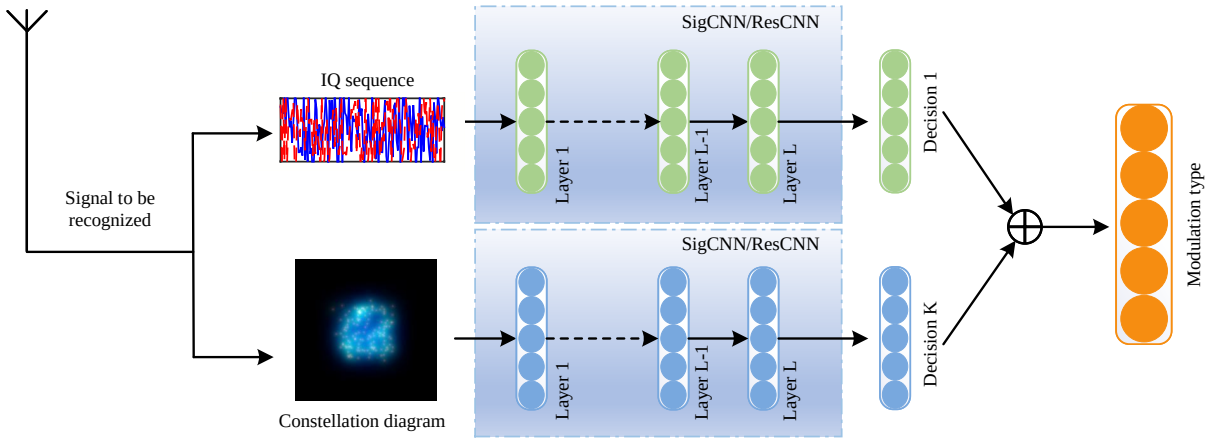


Figure 9. Proposed algorithm with late fusion.

Late fusion is a decision-level fusion method that aggregates the decision results of multiple classifiers, each of which is evaluated on different data modalities. The major challenge of late fusion is how decisions from different classifiers are combined.

Various fusion approaches, such as weighted averaging, learned meta-classifiers, majority voting, entropy-based weighting, confidence calibration, and Bayesian model averaging, can be applied in late fusion. For simplicity, this paper takes confidence-based approach as an example. This method effectively reduces the sensitivity of any single model to specific noise or anomalous inputs. First, the modulated signals are recognized from different signal modalities using a network model. The decision probability for the modulation type recognized by the k-th $(1 \leq k \leq K)$ signal modality is defined as $\mathbf{p^k}$. Thus, the decision probability vector for modulation recognition by the k-th signal modality can be written as

$$\mathbf{p^k} = \left[p_1^k, p_2^k, \ldots, p_M^k\right], \tag{16}$$

where

$$p_m^k = \frac{e^{G_L^k(\mathbf{f_m^k})}}{\sum_{i=1}^{M} e^{G_L^k(\mathbf{f_i^k})}}. \tag{17}$$

It should be pointed out that p_m^k represents the probability of recognizing the m types of modulation for the k-th modality, $\mathbf{f_m^k}$ represents the feature vector of the m types of modulation for the k-th modality, and $G_L^k(\cdot)$ represents the transfer function of the L layer. The decision probability vector for late fusion is

$$\mathbf{p^{late}} = \sum_{k=1}^{K} \mathbf{p^k}. \tag{18}$$

Finally, the index m corresponding to the maximum value in $\mathbf{p^{late}}$ is searched for, and it serves as the late fusion result for AMR

$$m = \arg\max \mathbf{p^{late}}. \tag{19}$$

Note that the late fusion method does not care about which signal modality and network model are utilized, but only focuses on the their results, namely the decision probability vector as shown in Equation (18). Consequently, it can be applied in various scenarios and has the best flexibility.

5. Experimental Results

This section provides experimental results to evaluate the performance of the proposed multimodal DL based AMR algorithms. For each modulation and signal modality, 4200 and 1800 observations are considered to form the training and test datasets. All training and test samples are generated independently. For AWGN and Rayleigh channels, each sample is simulated separately at each SNR, and the channel is regenerated for each modulation type and SNR to guarantee sample independence. The signal-to-noise ratio (SNR) of the modulated signal ranges from 0 dB to 10 dB in steps of 2 dB. In the experiments, the Adam optimizer is chosen, with a batch size set to 32 and a learning rate set to 0.001. The training process is conducted using keras and accelerated with the Nvidia RTX 4080. The training models are evaluated after up to 20,000 training iterations. The trained model with the highest validation accuracy is selected for later testing, while the validation set is independently generated and not used for iterative model parameter updating.

5.1. Recognition Accuracy

The recognition accuracy of the SigCNN model versus SNR under AWGN channel with the sample number of 128 is plotted in Figure 10. Five different algorithms, including multimodal DL based AMR with early, intermediate and late fusion as well as traditional AMR with single signal modality of either IQ sequence or constellation, are involved for comparison. According to this figure, the recognition accuracy of all algorithms increases as SNR increases. Moreover, it can be seen from this figure that both traditional algorithms with single signal modality exhibit inferior performance as expected. Compared with them, multimodal DL based algorithms exhibit higher recognition accuracy. It is because they utilize both signal modalities of IQ sequence and constellation diagram, no matter which fusion method is applied. Among them, the accuracy of intermediate fusion is the best, and that of late fusion is worst. These phenomena can be explained as follows. Intermediate fusion, which directly fuses the features of different signal modalities, is able to fully utilize the multimodal information of the modulated signal. Late fusion fuses the decision of multiple signal modalities, and some key information may have been lost after decision.

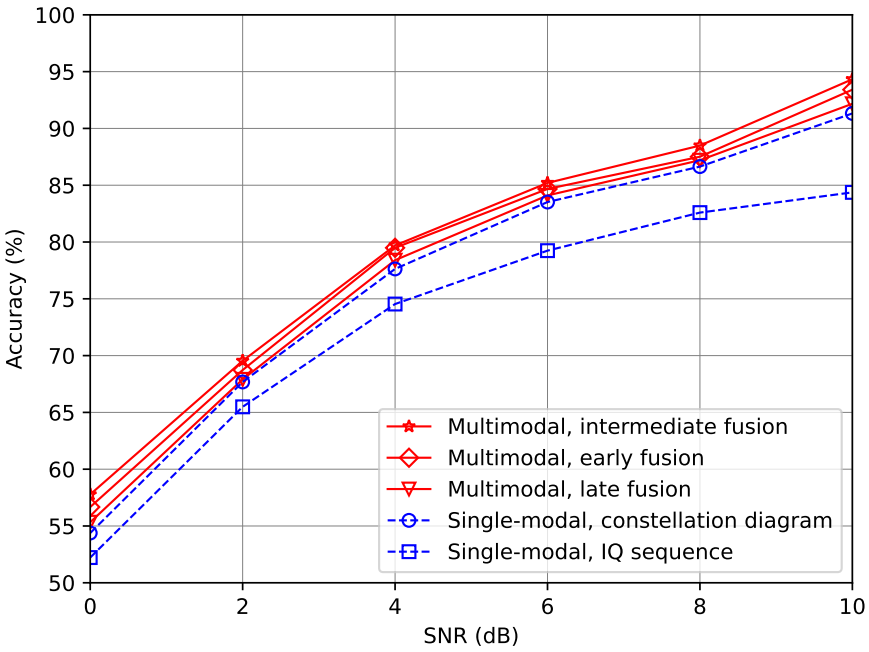


Figure 10. Recognition accuracy of the SigCNN model versus SNR with 128 samples.

The recognition accuracy of the ResCNN model versus SNR with 128 samples is depicted in Figure 11. As shown in this figure, although another network model is exploited, multimodal DL based algorithms similarly outperform traditional single signal modality based algorithms, and the algorithm that intermediately fuses IQ sequence and constellation diagram exhibits the highest recognition accuracy. Furthermore, compared with the SigCNN model, the accuracy of the ResCNN model is a bit better in high SNR region but a bit worse in low SNR region.

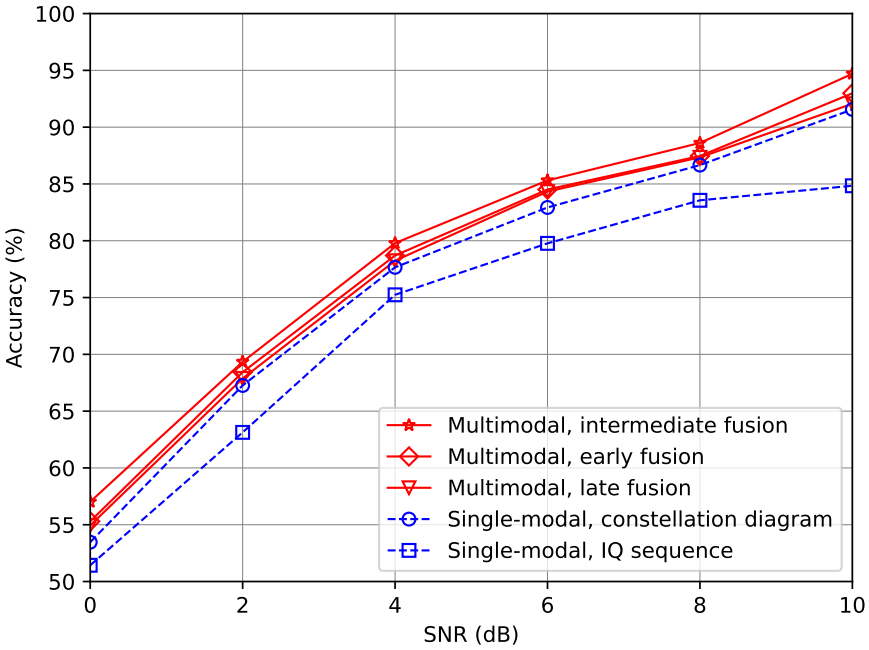


Figure 11. Recognition accuracy of the ResCNN model versus SNR with 128 samples.

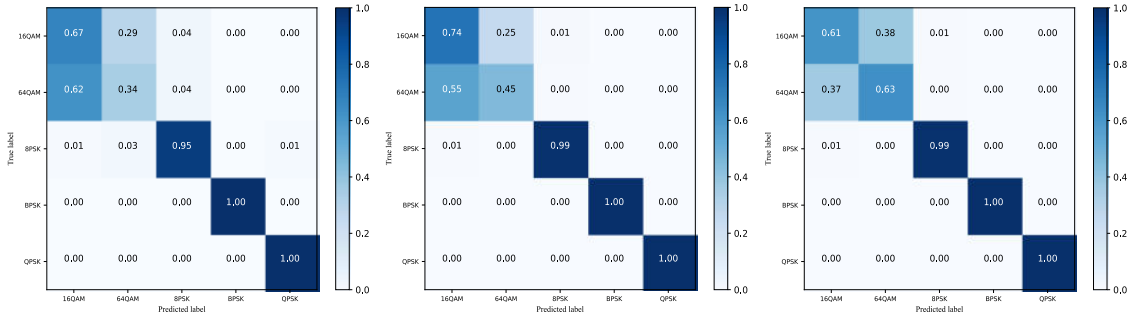
5.2. Confusion Matrix

The confusion matrix of the SigCNN model with 128 samples at the SNR of 6 dB is displayed in Figure 12. As depicted in this figure, no matter which algorithm is applied, BPSK and QPSK can be perfectly recognized with the recognition accuracy of 100%. It is because these modulation types have low modulation orders. When it comes to 8PSK, the recognition accuracy slightly decreases as the modulation order increases. Unfortunately, 16QAM and 64QAM are usually confused with each other, which cause numerous recognition errors. For the recognition of 64QAM, traditional single signal modality based algorithms obtain the accuracy of 34% and 45% when the signal modalities of IQ sequence and constellation diagram are separately utilized, respectively. As expected, the proposed multimodal DL based algorithms significantly improve the recognition accuracy of 64QAM to over 60% due to the joint utilization of two signal modalities. Particularly in the proposed algorithm with intermediate fusion, the recognition accuracy of 64QAM achieves 71%. This performance improvement benefits from the complementary advantages of multimodal features: the phase and temporal information carried by IQ sequences and the symbol distribution characteristics presented in constellation diagrams, helping the model better distinguish high-order QAM modulations.

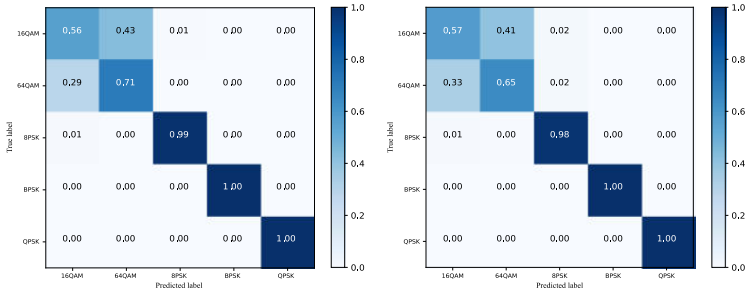
The F1-score of the SigCNN model with 128 samples at the SNR of 6 dB is described in Table 1. In the single-modal methods, the constellation diagram achieves the better performance, with F1-scores of 0.643 and 0.529 for 16QAM and 64QAM, respectively. These values significantly outperforming that of IQ sequence, which are 0.582 and 0.410, respectively. All three fusion methods achieve clear improvements in high-order modulation recognition. Among them, intermediate fusion performs the best, with the F1-score of 64QAM increasing to 0.663. Meanwhile, the F1-scores for 8PSK, BPSK, and QPSK all approach 1, resulting in an average F1-score of 0.851. The average F1-score of early fusion and late fusion is slightly lower than that of Intermediate fusion, but higher than that of single-modal. These results indicate that multimodal fusion can effectively alleviate the recognition bottleneck of high-order modulations, and intermediate fusion delivers the greatest advantage under the SigCNN model.

The confusion matrix of the ResCNN model with 128 samples at the SNR of 6 dB is described in Figure 13. According to this figure, despite the use of a different network model, especially the algorithm with intermediate fusion, still outperform single signal modality based algorithms in differentiating 64QAM that is high-order modulated and difficult to be recognized. Therefore, the superiority of applying multimodal DL in AMR is verified again.

The F1-score of the ResCNN model with 128 samples at the SNR of 6 dB is described in Table 2. Similarly, intermediate fusion reaches the highest average F1-score of 0.853, followed by early fusion at 0.846 and late fuison at 0.841. Moreover, all fusion methods are better than single-modal methods, in which the average F1-scores of constellation diagram and IQ sequence are 0.828 and 0.792, respectively. Consequently, the superiority of fusion methods, especially intermediate fusion, are verfifed again based on the ResCNN model.

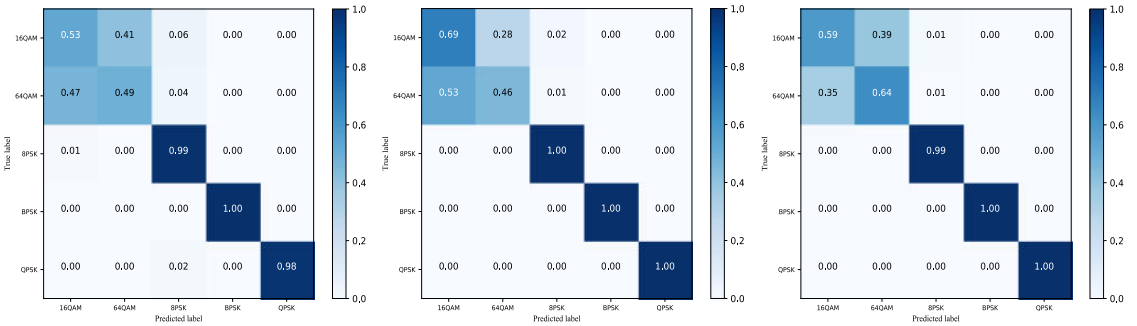


(a) IQ sequence (b) Constellation diagram (c) Early fusion

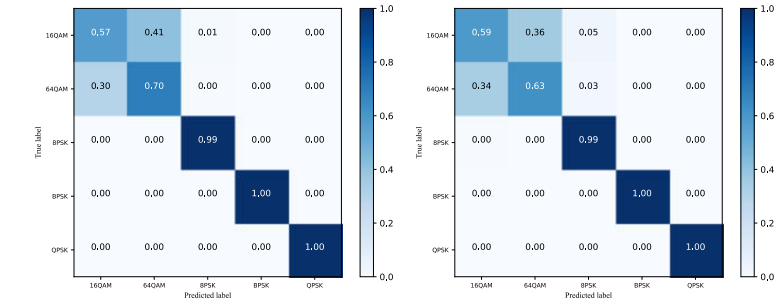


(d) Intermediate fusion (e) Late fusion

Figure 12. Confusion matrices of different algorithms based on the SigCNN model with 128 samples at the SNR of 6 dB: (**a**) IQ sequence; (**b**) Constellation diagram; (**c**) Early fusion; (**d**) Intermediate fusion; (**e**) Late fusion.



(a) IQ sequence (b) Constellation diagram (c) Early fusion



(d) Intermediate fusion (e) Late fusion

Figure 13. Confusion matrices of different algorithms based on the ResCNN model with 128 samples at the SNR of 6 dB: (**a**) IQ sequence; (**b**) Constellation diagram; (**c**) Early fusion; (**d**) Intermediate fusion; (**e**) Late fusion.

Table 1. F1-score of different algorithms based on the SigCNN model with 128 samples at the SNR of 6 dB.

Methods		16QAM	64QAM	8PSK	BPSK	QPSK	Average F1-Score
Single-modal	IQ sequence	0.582	0.410	0.936	1	0.995	0.785
	Constellation diagram	0.643	0.529	0.990	1	1	0.832
Multimodal	Early fusion	0.613	0.627	0.990	1	1	0.846
	Intermediate fusion	0.603	0.663	0.990	1	1	0.851
	Late fusion	0.597	0.630	0.975	1	1	0.840

Table 2. F1-score of different algorithms based on the ResCNN model with 128 samples at the SNR of 6 dB.

Methods		16QAM	64QAM	8PSK	BPSK	QPSK	Average F1-Score
Single-modal	IQ sequence	0.527	0.515	0.938	1	0.980	0.792
	Constellation diagram	0.625	0.529	0.985	1	1	0.828
Multimodal	Early fusion	0.611	0.630	0.990	1	1	0.846
	Intermediate fusion	0.613	0.664	0.990	1	1	0.853
	Late fusion	0.611	0.633	0.961	1	1	0.841

5.3. Impact of Sample Number

In order to highlight the performance of the proposed algorithms, we investigate the impact of sample number to the recognition accuracy. Figure 14 shows the recognition accuracy of different algorithms with the SigCNN model and 2 dB SNR at the sample number of 128 and 512. As shown in this figure, it is obvious that as the number of samples increases, all algorithms demonstrate superior recognition accuracy. Given a fixed sample number, compared the traditional single signal modality based algorithms, the multimodal DL based algorithms are able to improve the recognition accuracy. Moreover, the improvement in recognition accuracy is more significant when the sample number increases from 128 to 512. In the words, the multimodal DL based algorithms are more effective when fusing multiple signal modalities of the modulated signal with more samples.

The recognition accuracy of the ResCNN model versus the number of samples at the SNR of 2 dB is illustrated in Figure 15. Similar to Figure 14, multimodal DL based algorithms always outperforms traditional single signal modality based algorithms, and the superiority is more obvious when the sample number is larger.

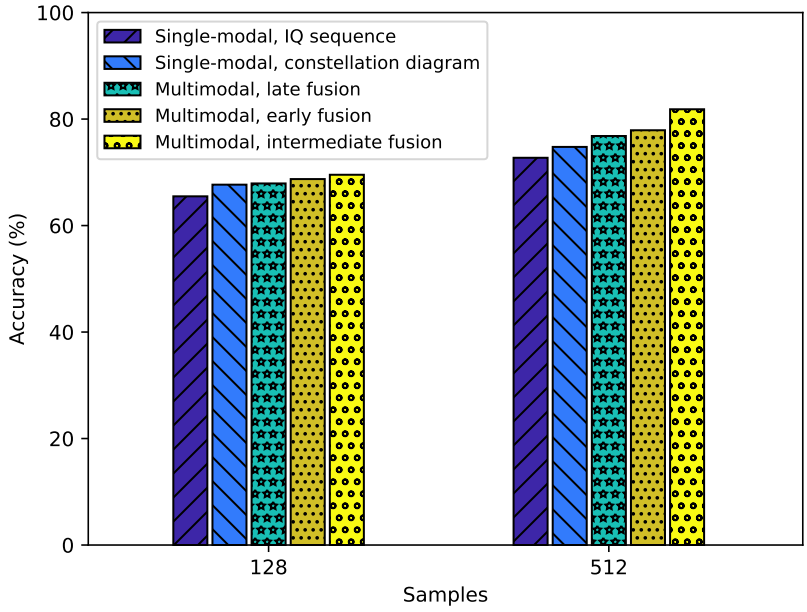


Figure 14. Recognition accuracy of different algorithms with the SigCNN model and 2 dB SNR at different samples.

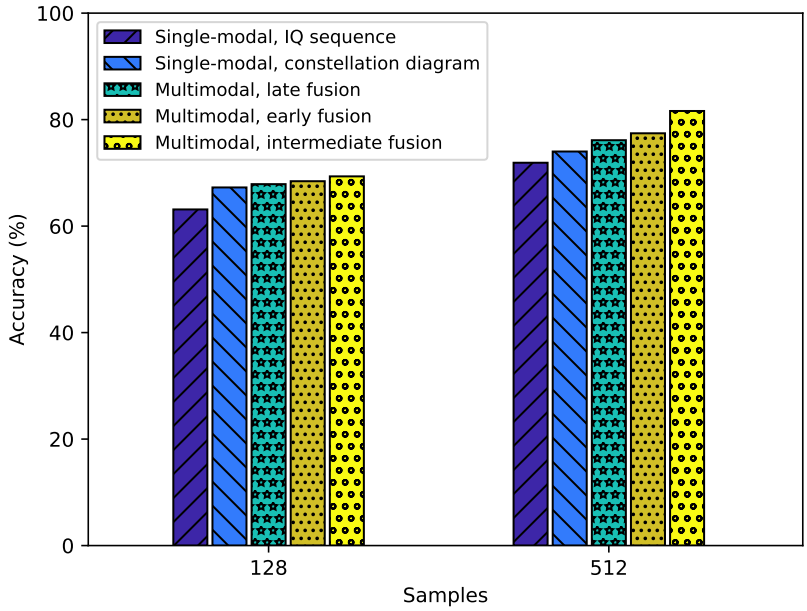


Figure 15. Recognition accuracy of different algorithms with the ResCNN model and 2 dB SNR at different samples.

5.4. Impact of Fading Channel

In order to evaluate the performance of the proposed algorithms under various scenarios, we consider that the modulated signal is transmitted through a Rayleigh fading channel, whose path delays and average path gains are set at $[0\ 500\ 800\ 1400\ 2000] \times 10^{-5}$ and $[0\ -5\ -10\ -15\ -20]$, respectively. Figure 16 depicts the recognition accuracy of the SigCNN model versus SNR with 128 samples under the Rayleigh fading channel. Compared with that under AWGN channel in Figure 10, the accuracy of all algorithms deteriorates severely due to the impact of fading channel. And Similar to Figure 10, no matter which fusion method is adopted, multimodal DL based algorithms are certainly better than the traditional algorithm with a single signal modality of either IQ sequence or constellation diagram under the tested experimental conditions. Moreover, the accuracy gaps between different algorithms are more significant under the fading channel, which indicates that the multimodal DL based algorithms are more powerful to combat the impact of fading channel and achieve accurate AMR. Furthermore, it can be observed from this figure that multimodal DL based algorithms achieve a greater degree of performance improvement than in AWGN for modulation recognition accuracy under a Rayleigh fading channel.

The recognition accuracy of the ResCNN model versus SNR with 128 samples under the Rayleigh fading channel is presented in Figure 17. As shown in this figure, despite the use of a different network model, multimodal DL based algorithms likewise outperform traditional single signal modality based algorithms, and the accuracy improvement under fading channel is greater than that under AWGN channel.

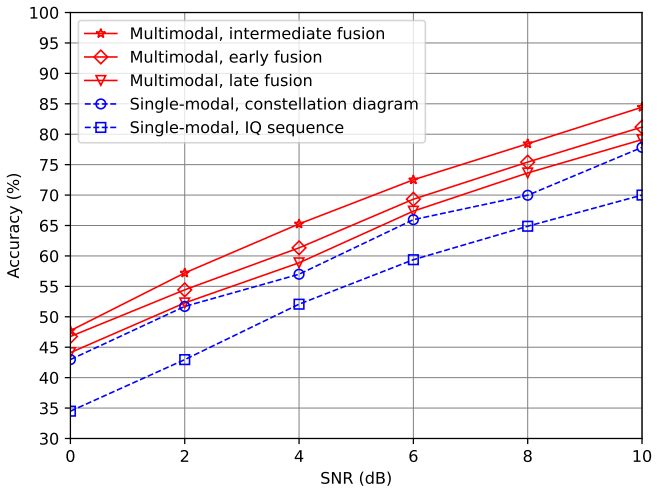


Figure 16. Recognition accuracy of the SigCNN model versus SNR with 128 samples under the Rayleigh fading channel.

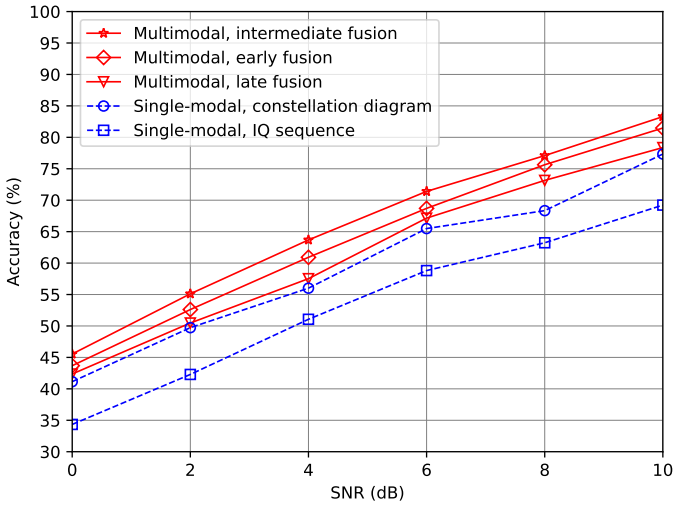


Figure 17. Recognition accuracy of the ResCNN model versus SNR with 128 samples under the Rayleigh fading channel.

5.5. Complexity and Feasibility

This subsection provides some results to demonstrate the complexity as well as feasibility of multimodal DL based algorithms. For fairness, all algorithms utilize the same sample number of 128. Table 3 shows the total number of model parameters used in different algorithms. According to this table, no matter which network model is applied, multimodal DL based algorithms consume more parameters compared with traditional single signal modality based algorithms. Detailedly, for early fusion, the increase in parameters is slight. It is because the algorithm with early fusion does not significantly modify the network structure but only add some neurons in the input layer. Unfortunately, the parameter numbers of algorithms with intermediate fusion and late fusion are nearly doubled. It may be resulted from the fact that these two algorithms use two branches of neural networks to extract the features of two signal modalities separately. The difference between these algorithms lies in that, in intermediate fusion, the extract features of two branches are firstly fused and then the fused features are utilized to make a decision, while in late fusion, the two-branch features are firstly utilized to make two individual decisions and then the individual decisions are fused to produce the final decision. Since one more decision-maker is used, the algorithm with late fusion consumes a bit more parameters than that with intermediate fusion as shown in Table 3.

Table 4 records the inference time of different algorithms with respect to two network models of the SigCNN and ResCNN. The inference time is averaged over 9000 tests excluding the preprocessing time for generating constellation diagrams. Similar to Table 3, all multimodal DL based algorithms have more inference time than traditional single signal modality based algorithms. In other words, the former achieves higher recognition accuracy as the cost of more calculation complexity. Among three multimodal DL based algorithms, early fusion and late fusion have the least and the most inference time, which implies that they have the lowest and the highest calculation complexity, respectively. Moreover, the inference time of all algorithms is at the microsecond level, which is acceptable for many practical communications systems. It is worth pointing out that Table 4 is concluded based on our experimental setup. With the advancement of DL and the evolution of calculation hardware, the feasibility of our algorithms will be further enhanced.

Table 5 quantitatively compares the influence of two factors, namely signal modality and network model on the three fusion methods, so as to evaluate their flexibility. In the table, "Yes" indicates that the method is affected by the factor, while "No" indicates that it is not affected. The results show that early fusion is affected by both signal modality and network model. This implies that extensive modifications or even redesign are required when the signal modality format or network model changes, resulting in the lowest flexibility. Intermediate fusion is only affected by the network model. It only needs to adapt to the feature map of the network's intermediate layers, without modifying the signal modality, thus exhibiting moderate flexibility. Late fusion is not affected by either of the two factors. Each modality is processed through an independent branch, so adding new modalities or replacing the backbone network does not require changes to the fusion method. Therefore, it incurs the least adaptation cost and achieves the highest flexibility.

Table 3. Parameters of different algorithms using the SigCNN and ResCNN.

Model	Parameters				
	IQ Sequence	Constellation Diagram	Early Fusion	Intermediate Fusion	Late Fusion
SigCNN	527,477	527,733	527,861	1,060,965	1,060,970
ResCNN	324,341	324,917	325,205	649,253	649,258

Table 4. Average inference time of different algorithms using the SigCNN and ResCNN.

Model	Inference Time (us)				
	IQ Sequence	Constellation Diagram	Early Fusion	Intermediate Fusion	Late Fusion
SigCNN	33.0	33.7	35.7	54.8	58.0
ResCNN	55.3	57.8	60.7	92.4	95.9

Table 5. Influencing factors.

	Early Fusion	Intermediate Fusion	Late Fusion
Signal modality	Yes	No	No
Network model	Yes	Yes	No
Number of factors needed to be adapted	2	1	0

6. Summary and Discussion

In this paper, we investigate three fusion methods, including early fusion, intermediate fusion and late fusion, for the multimodal DL based AMR. This section summarized the advantages as well as drawbacks of different methods from the perspectives of recognition accuracy, calculation complexity and application flexibility, as shown in Table 6. The performance metric of flexibility refers to the capability of a DL based AMR algorithm to adapt to new signal modalities or network structures. Since various promising modalities and networks have been emerging, flexibility is crucial to guarantee the superiority of the AMR algorithm.

Table 6. Summary of different fusion methods.

	Early Fusion	Intermediate Fusion	Late Fusion
Accuracy	✓	✓✓	❍
Complexity	✓✓	❍	❍
Flexibility	❍	✓	✓✓
	✓✓ Excellent	✓ Good ❍ Poor	

From the perspective of accuracy, intermediate fusion generally exhibits the best performance no matter which network model, sample number and channel condition are considered as it directly fuses the features of multiple signal modalities and makes full use of multilmodal information. On the contrary, late fusion generally performs the worst as it fuses the decisions of multiple signal modalities, during the process of which some key information for AMR might be lost.

From the perspective of complexity, as quantitatively compared in Table 3 (number of parameters) and Table 4 (average inference time), early fusion consumes the least model parameters as well as inference time, because it utilizes only one unified network branch. In contrast, both intermediate fusion and late fusion exploit two separate network branches for the two modalities, leading to a nearly doubled computational complexity compared with the early fusion.

From the perspective of flexibility, late fusion does not care about which signal modality as well as network model is adopted, and has the highest flexibility. Intermediate fusion is closely related to the network model used for AMR, and has the moderate flexibility. Early fusion, which depends on both the signal modality and the network model, exhibits the poorest flexibility.

7. Conclusions

This paper investigates the issue of multimodal DL based AMR, and proposes three algorithms that fuses multiple signal modalities in the early, intermediate and late stages, respectively. The core innovation of this work lies in comprehensively discussing the application of three fusion methods in AMR, breaking the limitations of single-modal signal representation in traditional AMR methods. Compared with the traditional algorithms that utilize a single signal modality, the proposed algorithms tend to achieve better recognition accuracy under tested scenarios, including different network models, sample sizes, and channel conditions. Moreover, this paper addresses the advantages as well as drawbacks of three algorithms from the accuracy, complexity and flexibility points of view. For a specific application, a suitable fusion method for multiple signal modalities can be chosen according to the requirements of AMR system. It should be pointed out that, although three fusion methods are illustrated by taking SigCNN and ResCNN as examples, they could be similarly applied in other state-of-the-art AMR models [36], such as CGDNet, CLDNN, and MCLDNN, could be similarly applied via multimodal fusion.

Author Contributions

X.L.: methodology, investigation, validation, writing—original draft; Q.G.: conceptualization, methodology, validation, writing—original draft; X.Y.: methodology, writing—reviewing and editing; P.W.: conceptualization, methodology, writing—reviewing; S.P.: conceptualization, methodology, supervision, reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Natural Science Foundation of China (Grant No. 62161012, 61861019), China Scholarship Council (Grant No. 202207540001), Fujian Provincial Department of Science and Technology (Grant No. 2025H0009) for their financial supports.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The dataset utilized in this study is described within the manuscript.

Conflicts of Interest

Given the role as Associate Editor, S.P. had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Pateromichelakis, E.; Bulakci, O.; Peng, C.; et al. LAA as a Key Enabler in Slice-Aware 5G RAN: Challenges and Opportunities. *IEEE Commun. Stand. Mag.* **2018**, *2*, 29–35.
2. Ya, T.; Yun, L.; Haoran, Z.; et al. Large-Scale Real-World Radio Signal Recognition with Deep Learning. *Chin. J. Aeronaut.* **2022**, *35*, 35–48.
3. Zou, Y.; Yao, Y.D.; Zheng, B. Cooperative Relay Techniques for Cognitive Radio Systems: Spectrum Sensing and Secondary User Transmissions. *IEEE Commun. Mag.* **2012**, *50*, 98–103.
4. Xin, C.; Song, M. Analysis of the On-Demand Spectrum Access Architecture for CBRS Cognitive Radio Networks. *IEEE Trans. Wireless Commun.* **2019**, *19*, 970–978.
5. Wang, Y.; Wang, J.; Zhang, W.; et al. Deep Learning-Based Cooperative Automatic Modulation Classification Method for MIMO Systems. *IEEE Trans. Veh. Technol.* **2020**, *69*, 4575–4579.
6. Zhang, H.; Farzanullah, M.; Sediq, A.B.; et al. Supervised Contrastive Learning for Uncertainty-Aware Wireless Signal Analysis: A Case Study for Modulation Classification. *IEEE Trans. Commun.* **2026**, *74*, 7227–7240.

7. Wang, G.; Liu, Z.; Zhang, X.; et al. PID: A Parameter-Efficient Isolation Domain-Incremental Learning Framework for Signal Modulation Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2026**, *37*, 1449–1462.
8. Shi, J.; Hong, S.; Cai, C.; et al. Deep Learning-Based Automatic Modulation Recognition Method in the Presence of Phase Offset. *IEEE Access* **2020**, *8*, 42841–42847.
9. Tunze, G.B.; Huynh-The, T.; Lee, J.M.; et al. Sparsely Connected CNN for Efficient Automatic Modulation Recognition. *IEEE Trans. Veh. Technol.* **2020**, *69*, 15557–15568.
10. Li, L.; Huang, J.; Cheng, Q.; et al. Automatic Modulation Recognition: A Few-Shot Learning Method Based on the Capsule Network. *IEEE Wireless Commun. Lett.* **2020**, *10*, 474–477.
11. Li, R.; Li, L.; Yang, S.; et al. Robust Automated VHF Modulation Recognition Based on Deep Convolutional Neural Networks. *IEEE Commun. Lett.* **2018**, *22*, 946–949.
12. Huang, L.; Pan, W.; Zhang, Y.; et al. Data Augmentation for Deep Learning-Based Radio Modulation Classification. *IEEE Access* **2019**, *8*, 1498–1506.
13. Zhang, F.; Luo, C.; Xu, J.; et al. An Efficient Deep Learning Model for Automatic Modulation Recognition Based on Parameter Estimation and Transformation. *IEEE Commun. Lett.* **2021**, *25*, 3287–3290.
14. Chen, Y.; Dong, B.; Liu, C.; et al. Abandon Locality: Frame-Wise Embedding Aided Transformer for Automatic Modulation Recognition. *IEEE Commun. Lett.* **2022**, *27*, 327–331.
15. Ramachandram, D.; Taylor, G.W. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108.
16. Liang, M.; Li, Z.; Chen, T.; et al. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *12*, 928–937.
17. Valada, A.; Oliveira, G.L.; Brox, T.; et al. Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion. In Proceedings of the 2016 International Symposium on Experimental Robotics, Nagasaki, Japan, 3–8 October 2016; pp. 465–477.
18. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
19. Jain, A.; Singh, A.; Koppula, H.S.; et al. Recurrent Neural Networks for Driver Activity Anticipation via Sensory-Fusion Architecture. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 3118–3125.
20. Peng, S.; Sun, S.; Yao, Y.D. A Survey of Modulation Classification Using Deep Learning: Signal Representation and Data Preprocessing. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 7020–7038.
21. Khan, F.N.; Zhong, K.; Al-Arashi, W.H.; et al. Modulation Format Identification in Coherent Receivers Using Deep Machine Learning. *IEEE Photonics Technol. Lett.* **2016**, *28*, 1886–1889.
22. Huang, J.; Yi, A.; Liao, P.; et al. Modulation Format Classification of Probabilistically Shaped M-QAM Signals Based on Nonlinear Power Transformation. In Proceedings of the 2022 Asia Communications and Photonics Conference (ACP), Shenzhen, China, 5–8 November 2022; pp. 782–785.
23. Peng, S.; Jiang, H.; Wang, H.; et al. Modulation Classification Based on Signal Constellation Diagrams and Deep Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 718–727.
24. Lin, Y.; Tu, Y.; Dou, Z.; et al. Contour Stella Image and Deep Learning for Signal Recognition in the Physical Layer. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *7*, 34–46.
25. Wang, D.; Zhang, M.; Li, Z.; et al. Modulation Format Recognition and OSNR Estimation Using CNN-Based Deep Learning. *IEEE Photonics Technol. Lett.* **2017**, *29*, 1667–1670.
26. Wang, Y.; Liu, M.; Yang, J.; et al. Data-Driven Deep Learning for Automatic Modulation Recognition in Cognitive Radios. *IEEE Trans. Veh. Technol.* **2019**, *68*, 4074–4077.
27. Qi, P.; Zhou, X.; Zheng, S.; et al. Automatic Modulation Classification Based on Deep Residual Networks with Multimodal Information. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *7*, 21–33.
28. Changbo, H.; Guowei, L.; Lijie, H.; et al. Multimodal Feature Fusion Recognition of Modulated Signals Based on Image and Waveform Domain. In Proceedings of the 2020 7th International Conference on Dependable Systems and Their Applications (DSA), Xi'an, China, 28–29 November 2020; pp. 337–342.
29. Deng, W.; Wang, X.; Huang, Z.; et al. Modulation Classifier: A Few-Shot Learning Semi-Supervised Method Based on Multimodal Information and Domain Adversarial Network. *IEEE Commun. Lett.* **2022**, *27*, 576–580.
30. Zhou, H.; Zhou, Z.; Bai, J. Electromagnetic Signal Modulation Classification Based on Multimodal Features and Reinforcement Learning. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–7.
31. Zheng, S.; Zhou, X.; Zhang, L.; et al. Toward Next-Generation Signal Intelligence: A Hybrid Knowledge and Data-Driven Deep Learning Framework for Radio Signal Classification. *IEEE Trans. Cogn. Commun. Netw.* **2023**, *9*, 564–579.
32. Xu, J.; Luo, C.; Parr, G.; et al. A Spatiotemporal Multi-Channel Learning Framework for Automatic Modulation Recognition. *IEEE Wireless Commun. Lett.* **2020**, *9*, 1629–1632.

33. O'Shea, T.J.; Roy, T.; Clancy, T.C. Over-the-Air Deep Learning Based Radio Signal Classification. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 168–179.
34. O'Shea, T.; Hoydis, J. An Introduction to Deep Learning for the Physical Layer. *IEEE Trans. Cogn. Commun. Netw.* **2017**, *3*, 563–575.
35. He, K.; Zhang, X.; Ren, S.; et al. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Zhang, F.; Luo, C.; Xu, J.; et al. Deep Learning Based Automatic Modulation Recognition: Models, Datasets, and Challenges. *Digit. Signal Process.* **2022**, *129*, 103650.