electronics

MDPI

Article

AbFTNet: An Efficient Transformer Network with Alignment before Fusion for Multimodal Automatic Modulation Recognition

Meng Ning [1], Fan Zhou [1,*], Wei Wang [2], Shaoqiang Wang [3], Peiying Zhang [4,5] and Jian Wang [6]

[1] School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China; ningmeng@sylu.edu.cn
[2] National Key Laboratory of Electromagnetic Space Security, Jiaxing 314000, China; wei_wang@mail.xidian.edu.cn
[3] School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266000, China; wangshaoqiang@qut.edu.cn
[4] Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; zhangpeiying@upc.edu.cn
[5] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250013, China
[6] College of Science, China University of Petroleum (East China), Qingdao 266580, China; wangjiannl@upc.edu.cn
* Correspondence: zhoufan@sylu.edu.cn

Abstract: Multimodal automatic modulation recognition (MAMR) has emerged as a prominent research area. The effective fusion of features from different modalities is crucial for MAMR tasks. An effective multimodal fusion mechanism should maximize the extraction and integration of complementary information. Recently, fusion methods based on cross-modal attention have shown high performance. However, they overlook the differences in information intensity between different modalities, suffering from quadratic complexity. To this end, we propose an efficient *Alignment before Fusion Transformer Network* (AbFTNet) based on an in-phase quadrature (I/Q) and Fractional Fourier Transform (FRFT). Specifically, we first align and correlate the feature representations of different single modalities to achieve mutual information maximization. The single modality feature representations are obtained using the self-attention mechanism of the Transformer. Then, we design an efficient cross-modal aggregation promoting (CAP) module. By designing the aggregation center, we integrate two modalities to achieve the adaptive complementary learning of modal features. This operation bridges the gap in information intensity between different modalities, enabling fair interaction. To verify the effectiveness of the proposed methods, we conduct experiments on the RML2016.10a dataset. The experimental results show that multimodal fusion features significantly outperform single-modal features in classification accuracy across different signal-to-noise ratios (SNRs). Compared to other methods, AbFTNet achieves an average accuracy of 64.59%, with a 1.36% improvement over the TLDNN method, reaching the state of the art.

Keywords: multimodal automatic modulation recognition; alignment before fusion; efficient cross-modal aggregation promoting module; transformer

Citation: Ning, M.; Zhou, F.; Wang, W.; Wang, S.; Zhang, P.; Wang, J. AbFTNet: An Efficient Transformer Network with Alignment before Fusion for Multimodal Automatic Modulation Recognition. *Electronics* **2024**, *13*, 3725. https://doi.org/10.3390/electronics13183725

Academic Editor: Xianzhi Wang

Received: 13 August 2024
Revised: 9 September 2024
Accepted: 9 September 2024
Published: 20 September 2024

1. Introduction

Automatic modulation recognition (AMR) technology is an essential component of both cooperative and non-cooperative communication scenarios in future communication systems [1–3]. In cooperative communication scenarios, AMR assists IoT devices in adaptively understanding and utilizing different modulation schemes, ensuring compatibility between devices and reducing communication interference [2,4,5]. In non-cooperative communication scenarios, AMR serves as an important tool for accurately identifying and classifying enemy communication and radar signals, providing crucial intelligence for electronic warfare [3,6,7].

In recent years, with the advancement of communication systems, signal data from sources such as radio [4], satellite [8], or optical fibers [9] have become increasingly diverse. As a result, multimodal automatic modulation recognition (MAMR) has emerged as a popular research direction [6,10]. The objective of this task is to identify the modulation scheme of a signal from a multimodal sequence, which may include modalities such as an in-phase quadrature (I/Q) [2,6,9,11–13], Fourier transform (FT) [5], and others. In general, the information from different modalities in a multimodal sequence is complementary. For example, I/Q focuses on the phase and amplitude of the signal, while FT focuses on the time frequency. Integrating the features of these two signals provides multidimensional information for the more accurate differentiation of modulation schemes. Therefore, the efficient multimodal fusion is the key to this task.

Existing AMR approaches mostly rely on deep learning methods. The task has evolved from convolutional neural networks [11,14–16] to self-attention mechanisms [6,7]. For instance, O'Shea et al. [11] introduced the RML2016.10a dataset, which mimics real wireless communication systems, and constructed a CNN model comprising convolutional and fully connected layers for classifying I/Q signals. Considering the temporal dependencies of signals, ref. [14] proposed the CLDNN model, which combines CNN and LSTM. Recently, the self-attention mechanism has emerged as a significantly improved method for modeling sequences [7,17]. The self-attention mechanism enables them to capture dependencies between different parts of the input sequence, offering significant opportunities for exploration in the field of AMR [6]. However, most AMR methods only use the temporal or frequency unimodal information, ignoring the complementarity of the multimodal information. The complexity and diversity of modern communication environments require learning rich feature representations and semantic information from multimodal data [18] to flexibly adapt to various signal features.

Multimodal fusion is the key for MAMR. Existing fusion methods include but are not limited to concatenation-based [15], tensor-based [16], CNN-based [17], and attention-based [6,18,19] approaches. For example, [18] proposed a dual-stream Transformer network, integrating both time-domain and spatial-domain I/Q features, demonstrating the effectiveness of multimodal features. Recently, methods based on cross-modal fusion have been gaining popularity [6,19]. Cross-modal attention enhances the target modality by learning directional pairwise attention between different elements. However, employing pairwise cross-modal attention to fuse multimodal sequences is inefficient, with computational complexity depending on the length of single modality features, often exhibiting quadratic complexity. For instance, directly enhancing the FT modality with the entire I/Q modality sequence may introduce redundant information. Therefore, achieving a balance between the model's efficiency and performance is crucial for MAMR. Furthermore, different modal representations exhibit varying levels of information intensity, leading to noise when directly fusing the information of different densities at the same layer. For instance, the signal intensity of shallow feature representations is low level, while that of deep features is high level. It is more comprehensive to fuse complementary information between high-level features. Moreover, the information from different modalities is asynchronous, and learning aligned correlated features facilitates aggregating inter-modal information, maximizing the exchange of information.

To this end, we propose an efficient alignment before fusion Transformer network (AbFTNet) based on I/Q and the Fractional Fourier Transform (FRFT). And the FRFT provides flexibility in adjusting the perspective of time–frequency analysis by introducing a fractional alpha parameter [20,21]. As shown in Figure 1, for the I/Q and FRFT modalities, we divide the architecture into three parts: single-modal feature extraction, alignment, and fusion. The core idea of this architecture is to align first and then fuse, which is achieved by using a contrastive learning method that minimizes the distance between positive samples and maximizes the distance between negative samples. The central module is used for multimodal fusion, reducing its complexity from quadratic to linear. The innovation of this paper, as illustrated in Figure 1, lies in the "align before fuse" approach and the linear

multimodal fusion module. This method provides an efficient solution for multimodal automatic modulation recognition. A more detailed architecture of the proposed AbFTNet is illustrated in Figure 2. Specifically, we first utilize Transformer to separately achieve single-modal learning, obtaining the single-modal representations of I/Q and FRFT. Next, we maximize mutual information between the two modalities through alignment strategies. We associate positive samples at the same position in the same batch and negative samples at different positions, utilizing CPC Loss [22,23] to achieve contrastive learning of different modalities. Finally, we design an efficient cross-modal aggregation-promoting module. To overcome the differences in information intensity between different modalities at different layers, we require one modality to organize and condense its information before sharing it with another modality. We integrate the two modalities through a message center, ensuring that one modality must pass through this message center before interacting with the other modality, facilitating the complementary learning of modal features. This operation reduces the quadratic complexity introduced by cross-modal attention and enhances the adaptive capabilities of different information levels. To validate the effectiveness of the proposed method, we conduct experiments on the publicly available RML2016.10a dataset [11]. The experimental results demonstrate that the multimodal approach outperforms unimodal methods and achieves state-of-the-art performance, with an accuracy of 64.59%.

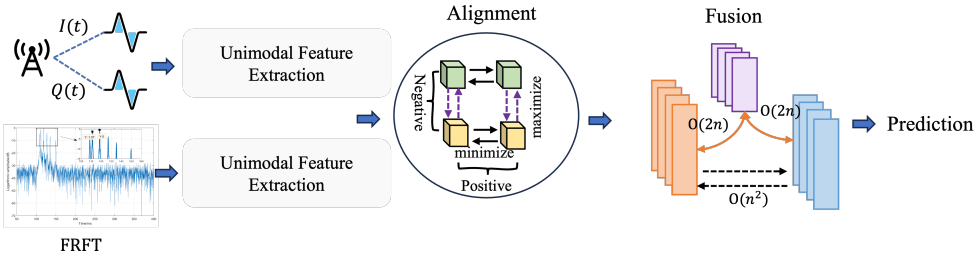


Figure 1. The streamlined architecture of the proposed AbFTNet.

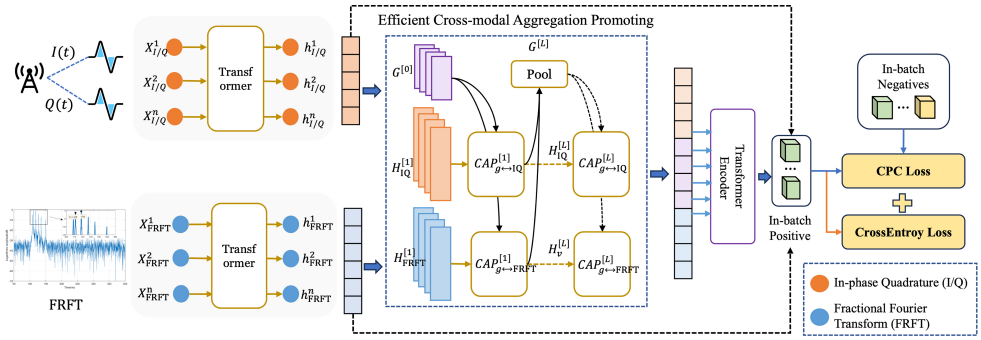


Figure 2. The architecture of the proposed AbFTNet. We propose an efficient Transformer network based on an in-phase quadrature (I/Q) and the Fractional Fourier Transform (FRFT) with alignment before fusion. We divide the architecture into three parts: unimodal feature extraction, alignment, and fusion. We design an efficient cross-modal aggregation-promoting module. This operation reduces the quadratic complexity introduced by cross-modal attention and enhances the adaptive capabilities of different information levels.

We summarize our **contributions** as follows:

- We propose an efficient alignment before fusion Transformer network (AbFTNet) based on I/Q and the Fractional Fourier Transform (FRFT).
- We design an efficient cross-modal aggregation-promoting (CAP) module. One modality organizes data before sharing, facilitated by a central message hub. This enhances complementary feature learning, reducing complexity and boosting adaptability across information levels.

- To validate the effectiveness of the proposed method, we conduct experiments on RML2016.10a. The results show that the AbFTNet outperforms unimodal methods, achieving state-of-the-art performance with an accuracy of 64.59%.

We summarize the **innovative** points of our work:

- For the multimodal automatic modulation recognition (MAMR) task, we propose a novel multimodal fusion network based on two modalities: I/Q and FRFT. The core concept of our network is to first align and then fuse the unimodal features. We achieve alignment through a contrastive learning method and fuse the features using a Transformer-based fusion network.This idea is used for the first time in the MAMR task.
- During the multimodal fusion phase, we introduce an efficient cross-modal aggregation module. While existing approaches often employ Cross-Attention for multimodal fusion, computing correlations between two modalities can be inefficient, with a complexity of $O(n^2)$ relative to the sequence length. To address this, we design a cross-modal aggregation module that improves efficiency. Each modality passes through this module before interacting with the other modality. Since the module length is only two, the time complexity of the two Cross-Attention operations is reduced to $O(2n + 2n) = O(n)$. This design enhances efficiency and significantly reduces the model's complexity.

The rest of this paper is organized as follows. Section 2 discusses relevant works on AMR and related topics. Section 3 provides a brief introduction to the wireless signal model. Section 4 elaborates on the structure and principles of the proposed framework. Section 5 presents the analysis of experimental results. Finally, Section 6 concludes the paper.

2. Related Work

Our proposed approach involves two main aspects: automatic modulation recognition, and automatic modulation recognition based on contrastive learning. The term "automatic modulation recognition" mainly includes unimodal extraction and multimodal fusion methods. This approach not only reduces complexity but also improves adaptability. Contrastive learning can establish effective representation learning between different signals. Next, we discuss related studies with respect to these two aspects.

2.1. Automatic Modulation Recognition

In communication systems, signals typically consist of lower-frequency components, whereas communication channels have limited bandwidth. Thus, modulation becomes essential to shift the signal to a higher frequency compatible with the channel before transmission. Accurate identification of the modulation type of the received signal is crucial during demodulation and decoding at the receiver end [24].

In recent years, deep learning has been widely applied in modulation recognition in the field of communication signals [3,25,26]. In order to simulate real wireless communication systems, the RML2016.10a dataset was established by [11], and a CNN model consisting of convolutional layers and fully connected layers was constructed for the classification of I/Q signals. Considering the temporal correlation of signals, [14] proposed the CLDNN model that combines CNN and LSTM. In [27], received signals were transformed into Angle Phase (AP) representation, and then AMR was performed using the LSTM model. Additionally, ref. [28] further applied GRU with fewer parameters and lower complexity, and proposed a gated residual recurrent neural network combined with ResNet. Although recurrent neural network (RNN) architectures such as LSTM and GRU can extract features from time series, their sequential structure suffers from problems like memory forgetting and inefficiency. In contrast, Transformer introduces the self-attention mechanism through parallel computation, which can simulate the correlation of global features. The MCFormer model based on Transformer was proposed in [6]. In this model, I/Q signals are treated as 1-D images with a depth of 2, and inputted into convolutional layers for embedding, where the size of convolutional kernels is k × 1. After pooling and flattening the output

of Transformer encoding blocks, encoded vectors can be obtained. In [2], I/Q sequences are divided into multiple equally long short sequences, and then modulation recognition is implemented in the form of images using Vision Transformer (ViT). In [5], an analysis of the accuracy of CNNs using I/Q, AP, and FT vectors was conducted, revealing that at a high signal-to-noise ratio (SNR), AP vectors achieved higher accuracy, while across all SNRs, I/Q vectors achieved higher average accuracy. Similarly, in [16], similar results were obtained for CGDNN and CLDNN models.

However, most AMR methods only use the temporal or frequency unimodal information, ignoring the complementary of the multimodal information [7,17,29]. Some researchers focus on cross-fusing different features to enhance their complementarity [30], or jointly utilizing modulation signal features [31]. In [7], a cascaded network comprising two CNNs was constructed, where I/Q sequences and constellation diagrams were separately utilized to train the two-level network for hierarchical modulation recognition. Experimental results demonstrate that the hierarchical recognition of QAMS and other signals through the two-level network resolves the confusion between 16QAM and 64QAM. Recently, methods based on cross-modal fusion have been gaining popularity [6,19]. Cross-modal attention enhances the target modality by learning directional pairwise attention between different elements. However, employing pairwise cross-modal attention to fuse multimodal sequences is inefficient, with computational complexity depending on the length of single modality features, often exhibiting quadratic complexity. In this paper, we design an efficient cross-modal aggregation-promoting module. This module reduces the quadratic complexity introduced by cross-modal attention and enhances the adaptive capabilities of different information levels.

2.2. Automatic Modulation Recognition Based on Contrastive Learning

In AMR, contrastive learning can assist the model in better understanding the characteristics of different modulation signals and establishing effective representation learning between different signals [19,32]. During training, the network often tends to rely more on modalities that are easy to extract information from, referred to as "strong modalities". At the same time, it tends to neglect other modalities that may contain other important information, termed "weak modalities". Consequently, the network becomes lazy and tends to converge to local optima, resulting in suboptimal outcomes, known as "lazy suboptimality". The key to contrastive learning lies in guiding the model to learn meaningful representations by comparing the similarity between different samples. In AMR, comparing samples of the same modulation type can bring signals of the same modulation type closer in the representation space, thereby enhancing the model's ability to distinguish similar signals. For instance, in [33], a limited number of labeled samples were utilized for pre-training. Pseudo-labels for unlabeled samples were then determined based on confidence. This method effectively augmented the dataset for retraining a customized CNN model. Ref. [19] proposed a Transformer-based contrastive learning semi-supervised learning framework. In [34], a semi-supervised signal recognition convolutional neural network (SSRCNN) with multiple loss functions was proposed. This method introduces Kullback–Leibler (KL) divergence and cross-entropy loss functions to handle unlabeled samples. However, when the number of labeled samples is small, the above method may not provide reliable pseudo-labels for unlabeled samples. In contrast, ref. [22] proposed a self-supervised method for learning representations from radio frequency signals. In [23], self-supervised contrastive learning (CL) was used for pre-training with labeled samples, constructing positive sample pairs through augmentation. After pre-training, the encoder parameters were fixed, and the classifier was trained using labeled samples. In this paper, we apply contrastive learning to AMR to enhance the robustness of the model.

3. Signal Model

The entire signal model can be represented by Figure 3. As shown in Figure 3, the generated symbol stream is described by $i(t)$, and the symbol sequence $i(t)$ is modulated into

a specific scheme to obtain the signal to be transmitted $s(t)$. Then, it is transmitted through a transmitting antenna. The signal $s(t)$ emitted by the transmitter can be represented as follows:

$$s(t) = S(i(t); \theta) \tag{1}$$

where S represents a function of the modulation process and θ represents the set of modulation parameters.

After undergoing channel propagation, the signal finally reaches the receiving antenna. Various impairments exist in the wireless channel during the propagation process, including but not limited to selective fading, propagation delay, and thermal noise. The received signal $r(t)$ is defined as follows:

$$r(t) = s(t) * h(t, \tau) + n(t) \tag{2}$$

where $h(t, \tau)$, τ, $n(t)$, and $*$ denote the channel impulse response function, channel parameters, additive Gaussian noise, and convolution, respectively.

In practical applications, the continuous signal $r(t)$ is sampled at a fixed sampling rate f_s and undergoes Hilbert transformation upon reception. Therefore, we can obtain the received time-domain discrete complex signal $r[n]$:

$$r[n] = r_I[n] + jr_Q[n] \tag{3}$$

where the real part and imaginary parts $r_I[n]$ and $r_Q[n]$ respectively represent the n-th I/Q component of the signal $r[n]$, with a sequence length of $N + 1$.

Then, the signal $r[n]$ is further represented as

$$r[n] = r_A[n]e^{j\pi r_P[n]} \tag{4}$$

where $r_A[n]$ and $r_P[n]$ respectively represent the n-th amplitude and phase components of the signal $r[n]$. At the receiver, the discrete signal $r[n]$ is processed through a modulation recognition module to determine the modulation type, and subsequently demodulated to extract the original information.

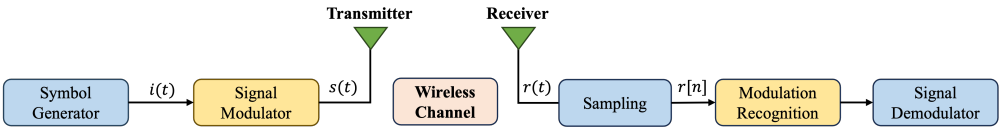


Figure 3. The entire signal model. At the transmitter, the symbol sequence $i(t)$ is modulated into a specific scheme and transmitted through a transmitting antenna. After channel propagation, the transmitted signal eventually reaches the receiving antenna. At the receiver, the discrete signal $r[n]$ is identified by the modulation recognition module to determine the modulation type, and then demodulated to retrieve the original information.

In automatic modulation recognition (AMR) based on deep learning, the raw discrete signal is commonly preprocessed into I/Q, A/P, TFI, or FT. In this paper, we directly use the I/Q representation to avoid extensive computation and achieve higher accuracy. The I/Q vectors are extracted from the orthogonal components of the discrete complex signal, combining to form a two-dimensional real-valued vector $X_{I/Q} \in \mathbb{R}^{(N+1)\times 2}$, expressed as

$$\mathbf{X}_{I/Q} = \begin{bmatrix} \mathbf{r}_I \\ \mathbf{r}_Q \end{bmatrix}^T = \begin{bmatrix} r_I[0] & r_I[1] & \cdots & r_I[N] \\ r_Q[0] & r_Q[1] & \cdots & r_Q[N] \end{bmatrix}^T \tag{5}$$

Furthermore, we find that the Fractional Fourier Transform (FRFT), by introducing a fractional order parameter to adjust the angle of time–frequency analysis, can be applied to a wider range of signal types. FRFT enables the examination of instantaneous behaviors of signals at different time and frequency points, including instantaneous frequency,

instantaneous amplitude, and instantaneous phase. These instantaneous statistics offer a more nuanced description of time–frequency characteristics, especially for signals with pronounced non-linearity and time-varying properties. When we input the complex signal $r[n]$, we obtain a one-dimensional vector $X_{FRFT} \in \mathbb{R}^{(N+1)\times 1}$ through the Fractional Fourier Transform, expressed as

$$r_{\emptyset}[\mu] = r[n]e^{-j\frac{2\pi k}{N+1}\alpha\mu}, (\mu \in [-N/2, N/2]) \tag{6}$$

Here, μ is a variable in the frequency domain, and $\emptyset$ is the fractional Fourier parameter, $\emptyset = 0.5\alpha\pi$. When $\alpha = 1$, the Fractional Fourier Transform is equivalent to the Fourier transform. The α is the adjustable hyperparameter. Thus, we have

$$\mathbf{X}_{FRFT} = [\mathbf{r}_{\emptyset}]^T = \begin{bmatrix} r_{\emptyset}[0] & r_{\emptyset}[1] & \cdots & r_{\emptyset}[N] \end{bmatrix}^T \tag{7}$$

4. Method

In this section, we propose innovative models for multimodal automatic modulation recognition that mainly involve two modalities, i.e., I/Q($\mathbf{X}_{I/Q}$) and FRFT($\mathbf{X}_{FRFT}$). We will develop and analyze our proposed AbFTNet model, in which, to efficiently interactively fuse the two modality sequences of the same signal, we aggregate intra- and inter-modal features, thereby achieving correct automatic modulation recognition. Specifically, we define a dataset D, consisting of n sequences $\{x_1, x_2, x_3, \ldots, x_n\}$ along with their corresponding labels $\{y_1, y_2, y_3, \ldots, y_n\}$. For the sequence data $x_m^i \in \mathbb{R}^{l_m \times d_m}, m \in \{I/Q, FRFT\}$, it includes its associated I/Q feature $\mathbf{X}_{I/Q}^i$ and FRFT feature $\mathbf{X}_{FRFT}^i$, l_m is the sequence length, and d_m represents the dimensionality of the representation vector for modality m, where $d_{I/Q} = 2, d_{FRFT} = 1$. We represent the multimodal automatic modulation recognition task as assigning a category y_i to each input x_i, where y_i indicates the corresponding true modulation category. The overall architecture is shown in Figure 2.

4.1. Unimodal Feature Extraction

We first encode the input multimodal sequence $\mathbf{X}_m, m \in \{\text{I/Q}, \text{FRFT}\}$. Specifically, we introduce two Transformer [35] layers to unify the features of each modality. We randomly initialize a low-dimensional set of tokens $H_m^0 \in \mathbb{R}^{l_m \times d_m}$ for each modality and use Transformer to embed the basic modality information into these tokens:

$$H_m^1 = \text{Transformer}(\text{Concat}(H_m^0, \mathbf{X}_m), \theta_m) \in \mathbb{R}^{T \times d} \tag{8}$$

where H_m^1 represents the single modality features for each modality, θ_m denotes the parameter information currently being learned by the Transformer, Concat($\bullet$) denotes the concatenation operation, T represents the length, and d represents the dimension. Due to the influence of the self-attention mechanism, T tokens can condense and integrate different single modality features.

Each Transformer consists of Multihead Self-Attention (MSA), Layer Normalization (LN), and Multilayer Perceptron (MLP) blocks applying residual connections. We denote a transformer layer, $h_m^{l+1} = \text{Transformer}(h_m^l)$, as

$$y^l = \text{MSA}(\text{LN}(h_m^l)) + h_m^l \quad m \in \{\text{I/Q}, \text{FRFT}\} \tag{9}$$

$$h_m^{l+1} = \text{MLP}(\text{LN}(y^l)) + y^l \quad m \in \{\text{I/Q}, \text{FRFT}\}. \tag{10}$$

Here, the MSA operation calculates the dot product attention, where the queries, keys and values are all linear projections of the same tensor h_m, $\text{MSA}(h_m) = \text{Attention}(W^Q h_m, W^K h_m, W^V h_m)$. This operation compresses information from each modality, sharing only essential data, maintaining multimodal fusion performance while reducing computational complexity. In practice, T and d are set to 8 and 128, respectively. The depth of the Transformer layer is set to 1. Additionally, transferring basic modality information to initialized

low-dimensional tokens helps reduce irrelevant redundant information, achieving higher efficiency with fewer parameters.

4.2. Multimodal Feature Fusion

In this section, we will perform multimodal fusion of I/Q and FRFT, enabling different modalities to acquire complementary and enhanced information. Recently, attention-based methods have mostly used Cross-Attention to achieve multimodal fusion, forming target modality reinforcement by learning directed pairwise attention between target and source modalities. Based on this, we designed an efficient cross-modal aggregation promoting (CAP). The architecture of CAP is illustrated in Figure 4. CAP takes two sequences, $X_{\mathrm{I/Q}}$ and X_{FRFT}, as input and outputs the mutually enhanced information $H_{\mathrm{I/Q}\to\mathrm{FRFT}}$ and $H_{\mathrm{FRFT}\to\mathrm{I/Q}}$. Specifically, the calculation of $\mathrm{CAP}_{\mathrm{I/Q}\to\mathrm{FRFT}}(X_{\mathrm{I/Q}}, X_{\mathrm{FRFT}})$ is as follows:

$$H'_{\mathrm{I/Q}\to\mathrm{FRFT}} = \mathrm{MCA}\left(\mathrm{LN}\left(X_{\mathrm{I/Q}}\right), LN(X_{\mathrm{FRFT}})\right) + X_{\mathrm{I/Q}} \tag{11}$$

$$H''_{\mathrm{I/Q}\to\mathrm{FRFT}} = \mathrm{MSA}\left(\mathrm{LN}\left(H'_{\mathrm{I/Q}\to\mathrm{FRFT}}\right)\right) + H'_{\mathrm{I/Q}\to\mathrm{FRFT}} \tag{12}$$

$$H_{\mathrm{I/Q}\to\mathrm{FRFT}} = \mathrm{MLP}\left(\mathrm{LN}\left(H''_{\mathrm{I/Q}\to\mathrm{FRFT}}\right)\right) + H''_{\mathrm{I/Q}\to\mathrm{FRFT}} \tag{13}$$

Similarly, we can obtain $\mathrm{CAP}_{\mathrm{FRFT}\to\mathrm{I/Q}}(X_{\mathrm{FRFT}}, X_{\mathrm{I/Q}})$. We define the cross-modal attention of two tensors X and Y, where X forms the query, and Y forms the keys and values used to reweight the query as $\mathrm{MCA}(X, Y) = \mathrm{Attention}(W^Q X, W^K Y, W^V Y)$.

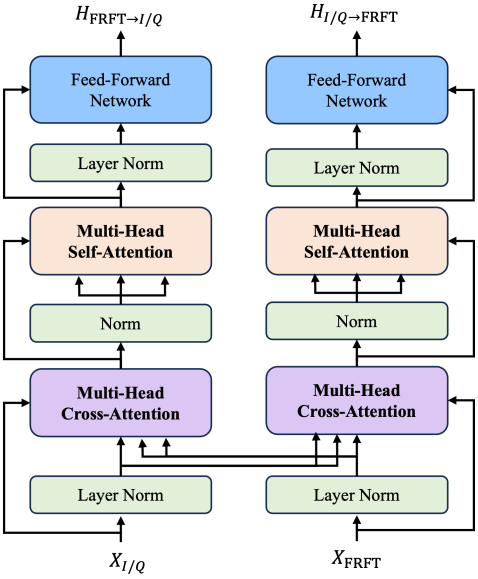


Figure 4. The architecture of the efficient cross-modal aggregation promoting (CAP). CAP takes two sequences, $X_{\mathrm{I/Q}}$ and X_{FRFT}, as input and outputs the mutually enhanced information $H_{\mathrm{I/Q}\to\mathrm{FRFT}}$ and $H_{\mathrm{FRFT}\to\mathrm{I/Q}}$.

4.3. Global-Local Interaction Learning Mode

We observed that the Cross-Attention is inefficient and introduces redundant features to the sequence. To further enhance modality interaction efficiency, we propose a global–local learning mode based on the linear computational cost. We believe that token-level representations from each modality can replace shared information and interact with local single-modality features as a global multimodal context G.

Specifically, we set the global multimodal context information $G[i] = \mathrm{concat}(h^i_{I/Q}, h^i_{FRFT}) \in \mathbb{R}^{2\times d}$, where i denotes the layer of global–local interaction. From the global information interacting with the local modality information, we learn modality consistency and specificity. The entire interaction process is as follows:

$$H_{I/Q}^{[i+1]}, G_{I/Q\to g}^{[i]} = CAP_{I/Q\longleftrightarrow g}^{[i]}(H_{I/Q}^{[i]}, G^{[i]}) \tag{14}$$

$$H_{FRFT}^{[i+1]}, G_{FRFT\to g}^{[i]} = CAP_{FRFT\longleftrightarrow g}^{[i]}(H_{FRFT}^{[i]}, G^{[i]}) \tag{15}$$

In this way, this strategy captures one-to-many global–local cross-modal interactions in two CAPs. By stacking multiple layers, the global multimodal context and local single-modality features can mutually reinforce and refine themselves. This operation requires M CAPs in each layer. Since the length of the global multimodal context is small, the overall time complexity reduces to $O\left(\sum_{m=1}^{M}(T_m + M)^2\right) \approx O\left(\sum_{m=1}^{M} T_m^2\right)$ (actually, $M \ll T_m$), degrading to $O(MT^2)$ in the case of modality alignment. Therefore, the default global–local fusion strategy in CAP not only has linear space complexity but also enjoys linear computation on involved modalities.

Additionally, each global–local fusion undergoes pooling layers to aggregate enhancement information from different modalities for subsequent fusion. We employ a fully connected layer with the tanh non-linear activation for this operation. We define two enhanced global multimodal contexts, $G_{FRFT\to g}^{[i]}$ and $G_{I/Q\to g}^{[i]}$. The new global multimodal context can be obtained as follows:

$$G^{[i+1]} = \text{softmax}(v^T \tanh(W^T G_g^T + b))G_g \tag{16}$$

where $G_g = \text{concat}(G_{I/Q\to g}^{[i]}, G_{I/Q\to g}^{[i]})$. We handle the entire learning process hierarchically, capturing different stage features of the model in each layer. We believe that the model learns shallow interaction features at the beginning and semantic features at later stages. This hierarchical learning method successfully integrates information from different modalities through cleverly designed aggregation blocks, providing a more comprehensive, rich multimodal feature representation for the model.

4.4. Contrastive Predictive Coding

CPC (Contrastive Predictive Coding) is a self-supervised learning method aimed at learning useful representations from data [36]. The objective of multimodal fusion is to project representations from different modalities into a common space and then aggregate modality-specific and modality-consistent information in that space. Therefore, alignment before multimodal fusion is crucial. CPC scores mutual information (MI) between context and future elements, compressing high-dimensional data into a more compact latent space, where conditional predictions are easier to model. Hence, we introduce contrastive learning.

For each anchor point, a batch of samples is randomly selected, consisting of two positive sample pairs and 2K negative sample pairs. Here, positive samples comprise modality pairs formed by the fused modality and the I/Q and corresponding FRFT within the same sample. Negative samples consist of modality pairs formed by the fused modality and the I/Q and FRFT from different samples. For each anchor point sample, the expression of self-supervised contrastive loss is as follows: the fused feature is defined as $\hat{X}_o$, and the features of the two modalities are defined as $h_{\text{I/Q}}$ and h_{FRFT}. A scoring function is applied to normalized prediction and ground truth vectors to measure their correlation:

$$G_{\varnothing}(\hat{X}_o) = \frac{G_z(\hat{X}_o)}{\|G_z(\hat{X}_o)\|_2} \tag{17}$$

$$h_m = \frac{h_m}{\|h_m\|_2} \tag{18}$$

$$s(h_m, \hat{X}_o) = exp\left(h_m(G_z(\hat{X}_o))^T\right) \tag{19}$$

Here, $G_{\varnothing}$ denotes a neural network consisting of parameters $\varnothing$ which computes the distance of fused features, and h_m represents the distance of modality features, where

$m \in \{\mathrm{I/Q}, \mathrm{FRFT}\}$. Adjusting how much information should be obtained from the fused modality is determined by incorporating this scoring function into the noise-contrastive estimation of similarity. Specifically,

$$\mathcal{L}_N^{\hat{X}_o, h_m} = -E[log \frac{s(h_m^i, \hat{X}_o)}{\sum s(h_m^j, \hat{X}_o)}] \tag{20}$$

$$\mathcal{L}_{CPC} = \mathcal{L}_N^{\hat{X}_o, h_{\mathrm{I/Q}}} + \mathcal{L}_N^{\hat{X}_o, h_{\mathrm{FRFT}}} \tag{21}$$

In summary, the entire multimodal model is optimized through two loss functions:

$$L = \mathcal{L}_{pred} + \alpha \mathcal{L}_{CPC} \tag{22}$$

where $\mathcal{L}_{pred}$ calculates the cross-entropy loss, and β represents the hyperparameter for adjusting the loss function. In the experiments, we set it to 0.1.

5. Experiments

In this section, we provide the details for experiments and results to demonstrate the performance and efficiency of our proposed AbFTNet. We mainly conduct experiments on the RML2016.10a dataset to verify the efficient performance. We perform comparative experiments, ablation experiments, and discuss them in terms of accuracy, parameters, FLOPs, and F1 scores, respectively.

5.1. Datasets

This paper employs the publicly available dataset RML2016.10a [11] in the field of modulation recognition as the research subject. This dataset encompasses 11 modulation signals under 20 signal-to-noise ratios (SNRs) ranging from −20 dB to + 18 dB with a step size of 2 dB. It includes 3 analog modulation signals, AM-DSB, AM-SSB, and WBFM, as well as 8 digital modulation signals, BPSK, QPSK, 8PSK, CPFSK, GFSK, PAM4, QAM16, and QAM64. Each modulation format at each SNR contains 1000 signal samples. Each sample comprises I/Q data, with each channel containing 128 samples. Additionally, to simulate real-world communication environments, the dataset incorporates various interference factors in the channel, such as the sampling frequency offset, center frequency offset, additive white Gaussian noise, multipath, and fading, providing challenging test conditions for research purposes.

5.2. Implementation Details

The experimental hyperparameters are shown in Table 1. The entire model is optimized using the Adam optimizer with a learning rate of 1×10^{-4}. Training and testing are conducted on a machine with an RTX 3090 GPU, using a batch size of 64. During the training process, we adaptively adjust the learning rate by reducing it by a factor of 10 when the validation accuracy no longer improves.

Table 1. The experimental hyperparameters setting.

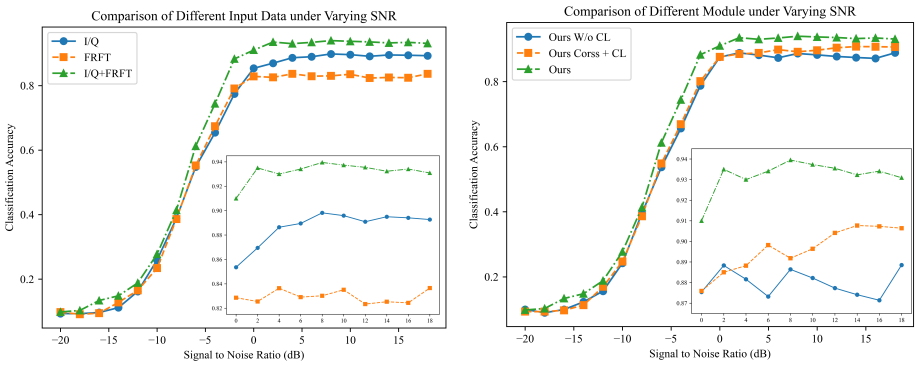
Hyperparameter	Value
T	8
d	128
Unimodal Depth	1
Multimodal Depth	3
Fusion Depth	2
β	0.1
Multihead Attention Dropout	0.4
LR	1×10^{-4}
Epochs	100

5.3. Ablation Study

We choose AbFTNet as the baseline network and validate the importance of multimodal features, contrast learning and multimodal fusion by conducting ablation experiments to observe the impact of different modules and unimodal features, such as AbFTNet without (W/O) contrast learning (CL), AbFTNet without (W/O) Cross-Attention (cross) and contrast learning (CL), AbFTNet with I/Q and AbFTNet with FRFT, etc.

5.3.1. The Effectiveness of Different Unimodal Modalities

To validate the effectiveness of multimodal features, we conduct separate experiments on single-modalities of I/Q and FRFT. As shown in Figure 5a, multimodal features outperform unimodal features overall. In the [0 dB–18 dB] range, the "I/Q + FRFT" multimodal performance exceeds that of unimodal features by 3% in accuracy. At higher signal-to-noise ratio levels, FRFT performs less effectively than I/Q overall. However, at lower signal-to-noise ratio levels, their performance is comparable. Nevertheless, multimodal features enhance unimodal performance to some extent, further capturing the complementary features of both, enhancing modulation recognition capabilities. At 0 dB, compared to I/Q and FRFT, respectively, there is a 6% and 7% improvement in accuracy.



(**a**) (**b**)

Figure 5. Assessing the effectiveness of different input modalities and exploring various modules in the ablation experiments of the RML2016.10a dataset. (**a**) The effectiveness of different input modalities. (**b**) The effectiveness of different modules.

5.3.2. The Effectiveness of Different Modules

For AbFTNet, we systematically remove different modules to verify their effectiveness. Firstly, we exclude the CAP module and utilize pairwise cross-modal attention for fusion ('Ours Cross + CL'). Figure 5b indicates a 4% accuracy decrease at 0 dB SNR and a 2–3% decrease at other SNR levels. This suggests that the designed CAP module not only considers complementary features from different modalities but further enhances the information brought by complementary information, maximizing the mutual information between modalities. Secondly, we solely retain CAP while removing CPC ('Ours W/o CL'). It can be observed that after removing contrastive learning, there is a significant decrease in the experimental results, with accuracy decreasing by 4–6%. It is worth noting that the performance of the experimental results is highly fluctuating at high signal-to-noise ratios, especially in the [8 dB–16 dB] range, where performance gradually declines. This indicates that fusion after alignment can effectively reduce noise interference caused by redundant information between modalities, enhancing robustness.

5.4. Comparison of AMR Methods

To evaluate the performance of our proposed method (AbFTNet), we compare it against other deep learning-based AMR methods. These methods include CLDNN, RESNET, CNN2, LSTM, DAE, MCLDNN, and TLDNN, among others. As shown in Figure 6, the AbFTNet model outperforms existing state-of-the-art (SOTA) methods across

all signal-to-noise ratio (SNR) levels on the Radio2016.10a dataset, achieving the highest average accuracy of 64.59% (+1.76%). In particular, AbFTNet exhibits significant performance in the mid-to-low SNR range, showing similar performance to TLDNN and notably surpassing other methods. AbFTNet achieves an accuracy of 59.6% at −6 dB, representing low-SNR conditions. In high-SNR scenarios, AbFTNet achieves the highest accuracy of 93.8% at 12 dB, while TLDNN reaches a maximum accuracy of 92.3%. This indicates that AbFTNet possesses a unique advantage in handling noise, demonstrating outstanding performance in low-SNR environments while maintaining generalizability across all noise environments. We analyze that, on one hand, compared to other methods, AbFTNet employs an alignment-first, fusion-later approach, which imposes an information entropy constraint on the features, maximizing the complementary information in multimodal fusion. On the other hand, since signals are often redundant, AbFTNet uses a more efficient fusion scheme, avoiding the impact of noise at low signal-to-noise ratios.

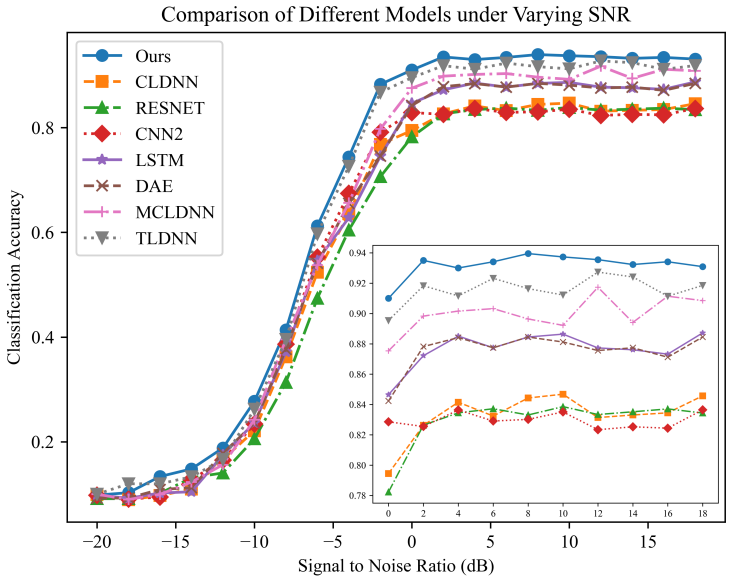


Figure 6. Comparison of Different Models on RML2016.10a Dataset.

5.5. Comparison of Computational Complexity

To further validate the efficiency of the proposed method, we compare the parameter and computational complexities of different models. The results are shown in Table 2, where KB (Kilobytes) refers to a unit of data storage equivalent to 1024 bytes. It is often used to quantify the size of model parameters in machine learning. MB (MegaFLOPS) refers to a measurement of computational performance, indicating millions of floating-point operations per second. From Table 2, we observe the following: (1) Compared to the low-parameter and low-computational-complexity DAE method, our method achieved an average accuracy improvement of 3.46%. (2) In terms of computational speed, AbFTNet outperformed the existing best models, MCLDNN and LSTM, achieving a substantial reduction of 80% to 90% in FLOPS. (3) Compared to TLDNN, our proposed method not only exhibited better performance but also higher accuracy, with an improvement of 1.76%. Overall, our proposed method further advances research development in balancing parameters and accuracy.

Table 2. Comparison of network computational complexity.

Model	Parameters (KB)	FLOPs (MB)	Runtime (ms)	Average Acc (%)
CLDNN [14]	25.35	3.70	103	57.02
RESNET [37]	241.81	11.49	323	57.76
CNN2 [38]	2749.28	19.06	535	58.54
LSTM [27]	201.10	25.82	725	60.33
PET-CGDNN [39]	71.87	4.6	130	60.44
DAE [40]	14.99	1.75	49	60.93
MCLDNN [41]	406.20	49.15	1381	61.47
CGFHNN [42]	319.42	72.95	2049	62.23
MobileAmcT [43]	303.47	46.06	1294	62.58
TLDNN [44]	243.34	7.89	221	62.83
Ours	175.69	6.96	195	64.59

5.6. Comparison of Different Modulation Classes

The F1 score is a metric for classification tasks that accurately evaluates the performance of the current model. Therefore, we compare the proposed method with other methods on the RML2016.10a dataset for 8PSK, AM-DSB, 64-QAM, and QPSK modulation schemes at 0 dB and 10 dB. Specifically, as shown in Figure 7, it can be seen that our method reaches a comparable level on these four modulation schemes, with improvements observed. Compared to TLDNN, our method shows the most significant improvements on these four modulation schemes, with increases of 5%, 5%, 8%, and 1% at 0 dB, respectively. It is worth noting that our method achieves a 100% F1 score for 8PSK and AM-DSB modulation schemes at both 0 dB and 10 dB, demonstrating outstanding performance.

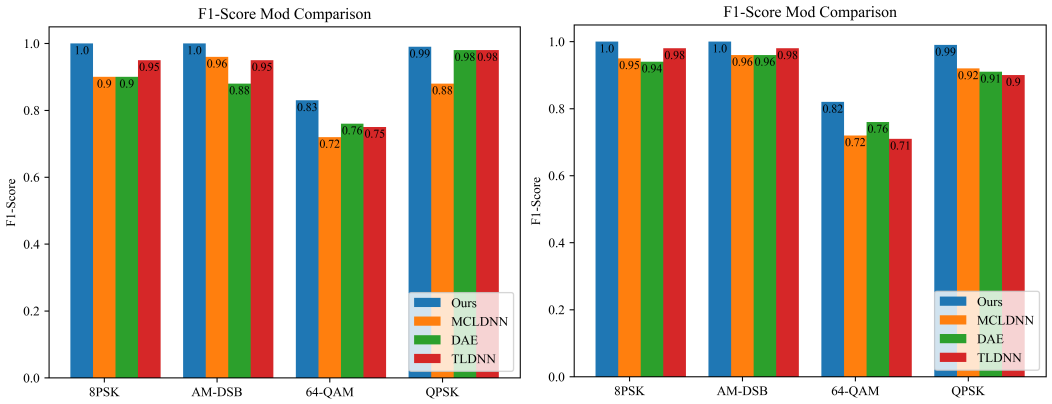


Figure 7. The F1 scores for the 8PSK, AM-DSB, 64-QAM, and QPSK modulation classes in the RML2016.10a dataset at 0 dB (**left**) and 10 dB (**right**) SNR.

The confusion matrix serves as a standard format for evaluating accuracy, providing further insight into the accuracy across different categories. Therefore, we compare our proposed method with other models on the RML2016.10a dataset, focusing on the confusion matrices for different modulation classes at the 0 dB and 10 dB SNR levels. The results are shown in Figure 8. In the 0 dB range, our method consistently outperforms MCLDNN. Particularly, significant accuracy improvements of nearly 14% are observed for the 16-QAM and 64-QAM modulation classes. In the 10 dB range, in addition to similar results to those at 0 dB, we achieve 100% accuracy in many modulation classes. This further demonstrates the robustness and effectiveness of the multimodal approach in feature representation and fusion.

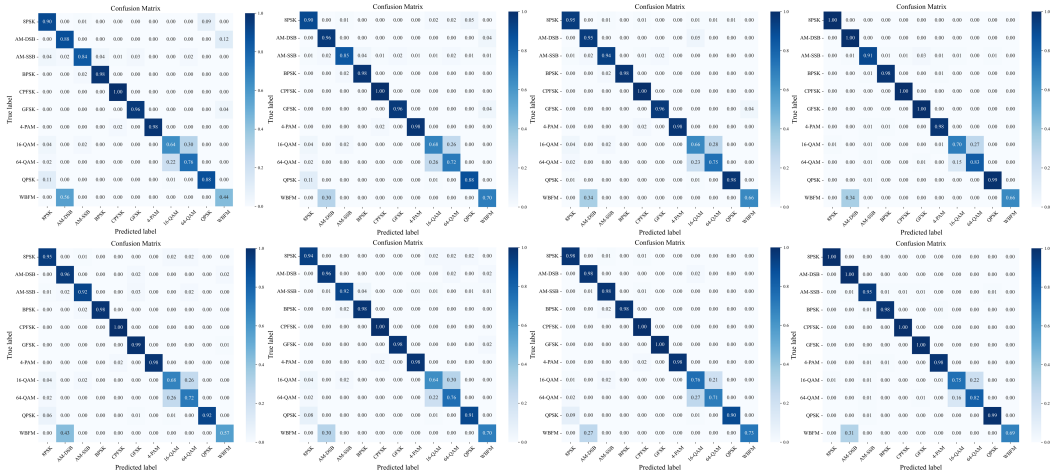


Figure 8. The proposed method and other models are compared based on the confusion matrices for different modulation classes at 0 dB and 10 dB SNR levels on the RML2016.10a dataset.

6. Conclusions

In this paper, we propose AbFTNet, an efficient Transformer network integrating I/Q and FRFT with alignment before fusion. Our method maximizes mutual information by aligning and correlating feature representations from different modalities. We introduce a cross-modal aggregation promoting (CAP) module to facilitate the adaptive complementary learning of modal features. This module serves as an integration hub for the modalities, fostering the adaptive complementary learning of features. By bridging the gap in information intensity between modalities, our approach ensures fair interaction and enhances overall performance. Experimental validation on the RML2016.10a dataset underscores the superiority of multimodal methods over unimodal ones. AbFTNet achieves an average accuracy of 64.59%, surpassing state-of-the-art benchmarks. This underscores the efficacy of our proposed approach in advancing the field of MAMR. This result not only improves the performance of MAMR but also provides a solution for real-time automatic modulation recognition. In future work, we will incorporate more modalities such as wavelet transforms, constellation diagrams, etc., and will validate on more datasets, both publicly available simulated datasets and real datasets.

Author Contributions: Conceptualization, M.N.; Methodology, F.Z.; Validation, M.N.; Formal analysis, M.N.; Supervision, W.W., S.W., P.Z. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by Shenyang Science and Technology Program (NO.23-503-6-16), General Project of Education Department of Liaoning Province in 2022 (NO.LJKMZ 2022 0612), Joint Fund of the Ministry of Education for Pre-Equipment Research 2023, the National Natural Science Foundation of China under Grant 62471493, the Shandong Provincial Natural Science Foundation under Grant ZR2023LZH017, ZR2022LZH015.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Huynh-The, T.; Pham, Q.V.; Nguyen, T.V.; Nguyen, T.T.; Ruby, R.; Zeng, M.; Kim, D.S. Automatic modulation classification: A deep architecture survey. *IEEE Access* **2021**, *9*, 142950–142971. [CrossRef]
2. Cai, J.; Gan, F.; Cao, X.; Liu, W. Signal modulation classification based on the transformer network. *IEEE Trans. Cogn. Commun. Netw.* **2022**, *8*, 1348–1357. [CrossRef]
3. Zhang, F.; Luo, C.; Xu, J.; Luo, Y.; Zheng, F.C. Deep learning based automatic modulation recognition: Models, datasets, and challenges. *Digit. Signal Process.* **2022**, *129*, 103650. [CrossRef]
4. Dobre, O.A.; Abdi, A.; Bar-Ness, Y.; Su, W. Survey of automatic modulation classification techniques: Classical approaches and new trends. *IET Commun.* **2007**, *1*, 137–156. [CrossRef]

5. Kulin, M.; Kazaz, T.; Moerman, I.; De Poorter, E. End-to-end learning from spectrum data: A deep learning approach for wireless signal identification in spectrum monitoring applications. *IEEE Access* **2018**, *6*, 18484–18501. [CrossRef]
6. Ma, W.; Cai, Z.; Wang, C. A Transformer and Convolution-Based Learning Framework for Automatic Modulation Classification. *IEEE Commun. Lett.* **2024**, *28*, 1392–1396. [CrossRef]
7. Zhai, L.; Li, Y.; Feng, Z.; Yang, S.; Tan, H. Learning Cross-Domain Features With Dual-Path Signal Transformer. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, 1–7. [CrossRef]
8. Zeng, Y.; Zhang, M.; Han, F.; Gong, Y.; Zhang, J. Spectrum analysis and convolutional neural network for automatic modulation recognition. *IEEE Wirel. Commun. Lett.* **2019**, *8*, 929–932. [CrossRef]
9. Sun, Z.; Wang, G.; Zhai, G.; Li, P.; Liang, Q.; Zhang, M. Signal detection and material identification method for loose particles inside sealed relays based on fusion classification model. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107284. [CrossRef]
10. Zhang, J.; Wang, T.; Feng, Z.; Yang, S. Towards the automatic modulation classification with adaptive wavelet network. *IEEE Trans. Cogn. Commun. Netw.* **2023**, *9*, 549–563. [CrossRef]
11. O'Shea, T.J.; Corgan, J.; Clancy, T.C. Convolutional radio modulation recognition networks. In Proceedings of the Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, 2–5 September 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 213–226.
12. Kong, W.; Yang, Q.; Jiao, X.; Niu, Y.; Ji, G. A transformer-based CTDNN structure for automatic modulation recognition. In Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 10–13 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 159–163.
13. Fei, S.; Zhang, C. Research on Modulation Mode Recognition Based on Dual-channel Hybrid Network Model. *J. Shenyang Ligong Univ.* **2023**, *42*, 34–39.
14. Ramjee, S.; Ju, S.; Yang, D.; Liu, X.; Gamal, A.E.; Eldar, Y.C. Fast deep learning for automatic modulation classification. *arXiv* **2019**, arXiv:1901.05850.
15. Chen, T.; Gao, S.; Zheng, S.; Yu, S.; Xuan, Q.; Lou, C.; Yang, X. Emd and vmd empowered deep learning for radio modulation recognition. *IEEE Trans. Cogn. Commun. Netw.* **2022**, *9*, 43–57. [CrossRef]
16. Chang, S.; Huang, S.; Zhang, R.; Feng, Z.; Liu, L. Multitask-learning-based deep neural network for automatic modulation classification. *IEEE Internet Things J.* **2021**, *9*, 2192–2206. [CrossRef]
17. Wang, Y.; Liu, M.; Yang, J.; Gui, G. Data-driven deep learning for automatic modulation recognition in cognitive radios. *IEEE Trans. Veh. Technol.* **2019**, *68*, 4074–4077. [CrossRef]
18. Qi, P.; Zhou, X.; Zheng, S.; Li, Z. Automatic modulation classification based on deep residual networks with multimodal information. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *7*, 21–33. [CrossRef]
19. Kong, W.; Jiao, X.; Xu, Y.; Zhang, B.; Yang, Q. A transformer-based contrastive semi-supervised learning framework for automatic modulation recognition. *IEEE Trans. Cogn. Commun. Netw.* **2023**, *9*, 950–962. [CrossRef]
20. Satija, U.; Mohanty, M.; Ramkumar, B. Automatic modulation classification using S-transform based features. In Proceedings of the 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 19–20 February 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 708–712.
21. Ren, B.; Teh, K.C.; An, H.; Gunawan, E. Automatic Modulation Recognition of Dual-Component Radar Signals Using ResSwinT-SwinT Network. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 6405–6418. [CrossRef]
22. Davaslioglu, K.; Boztaş, S.; Ertem, M.C.; Sagduyu, Y.E.; Ayanoglu, E. Self-supervised RF signal representation learning for NextG signal classification with deep learning. *IEEE Wirel. Commun. Lett.* **2022**, *12*, 65–69. [CrossRef]
23. Liu, D.; Wang, P.; Wang, T.; Abdelzaher, T. Self-contrastive learning based semi-supervised radio modulation classification. In Proceedings of the MILCOM 2021–2021 IEEE Military Communications Conference (MILCOM), San Diego, CA, USA, 29 November–2 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 777–782.
24. Zheng, S.; Chen, S.; Yang, X. DeepReceiver: A deep learning-based intelligent receiver for wireless communications in the physical layer. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *7*, 5–20. [CrossRef]
25. Kim, B.; Mecklenbräuker, C.; Gerstoft, P. Deep Learning-based Modulation Classification of Practical OFDM Signals for Spectrum Sensing. *arXiv* **2024**, arXiv:2403.19292.
26. Zhou, R.; Liu, F.; Gravelle, C.W. Deep learning for modulation recognition: A survey with a demonstration. *IEEE Access* **2020**, *8*, 67366–67376. [CrossRef]
27. Rajendran, S.; Meert, W.; Giustiniano, D.; Lenders, V.; Pollin, S. Deep learning models for wireless signal classification with distributed low-cost spectrum sensors. *IEEE Trans. Cogn. Commun. Netw.* **2018**, *4*, 433–445. [CrossRef]
28. Huang, S.; Dai, R.; Huang, J.; Yao, Y.; Gao, Y.; Ning, F.; Feng, Z. Automatic modulation classification using gated recurrent residual network. *IEEE Internet Things J.* **2020**, *7*, 7795–7807. [CrossRef]
29. Lin, S.; Zeng, Y.; Gong, Y. Learning of time-frequency attention mechanism for automatic modulation recognition. *IEEE Wirel. Commun. Lett.* **2022**, *11*, 707–711. [CrossRef]
30. Zhang, X.; Li, T.; Gong, P.; Liu, R.; Zha, X. Modulation recognition of communication signals based on multimodal feature fusion. *Sensors* **2022**, *22*, 6539. [CrossRef]
31. Deng, W.; Wang, X.; Huang, Z.; Xu, Q. Modulation classifier: A few-shot learning semi-supervised method based on multimodal information and domain adversarial network. *IEEE Commun. Lett.* **2022**, *27*, 576–580. [CrossRef]

32. Liu, F.; Pan, J.; Zhou, R. Contrastive learning-based multimodal fusion model for Automatic Modulation Recognition. *IEEE Commun. Lett.* **2023**, *28*, 78–82. [CrossRef]
33. Zhang, Y.; Zhao, Z. Limited data spectrum sensing based on semi-supervised deep neural network. *IEEE Access* **2021**, *9*, 166423–166435. [CrossRef]
34. Dong, Y.; Jiang, X.; Cheng, L.; Shi, Q. SSRCNN: A semi-supervised learning framework for signal recognition. *IEEE Trans. Cogn. Commun. Netw.* **2021**, *7*, 780–789. [CrossRef]
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
36. Gutmann, M.; Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 297–304.
37. 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. [CrossRef]
38. Chen, Y.; Dong, B.; Liu, C.; Xiong, W.; Li, S. Abandon locality: Frame-wise embedding aided transformer for automatic modulation recognition. *IEEE Commun. Lett.* **2022**, *27*, 327–331. [CrossRef]
39. Zhang, F.; Luo, C.; Xu, J.; Luo, Y. An efficient deep learning model for automatic modulation recognition based on parameter estimation and transformation. *IEEE Commun. Lett.* **2021**, *25*, 3287–3290. [CrossRef]
40. Ke, Z.; Vikalo, H. Real-time radio technology and modulation classification via an LSTM auto-encoder. *IEEE Trans. Wirel. Commun.* **2021**, *21*, 370–382. [CrossRef]
41. Xu, J.; Luo, C.; Parr, G.; Luo, Y. A spatiotemporal multi-channel learning framework for automatic modulation recognition. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 1629–1632. [CrossRef]
42. Luo, Q.; Zhao, M.M.; Chen, Z.; Su, Z.; Zhao, M.J. Complex-valued convolution and frequency global filter for automatic modulation recognition. *IEEE Commun. Lett.* **2023**, *27*, 1779–1783. [CrossRef]
43. Fei, H.; Wang, B.; Wang, H.; Fang, M.; Wang, N.; Ran, X.; Liu, Y.; Qi, M. MobileAmcT: A Lightweight Mobile Automatic Modulation Classification Transformer in Drone Communication Systems. *Drones* **2024**, *8*, 357. [CrossRef]
44. Qu, Y.; Lu, Z.; Zeng, R.; Wang, J.; Wang, J. Enhancing Automatic Modulation Recognition through Robust Global Feature Extraction. *arXiv* **2024**, arXiv:2401.01056.