IDAF: Iterative Dual-channel Attentional Fusion for Automatic Modulation Recognition

Bohan Liu , Ruixing Ge , Yuxuan Zhu , Bolin Zhang , Xiaokai Zhang , Yanfei Bao *

Posted Date: 6 September 2023

doi: 10.20944/preprints202309.0234.v1

Keywords: Multi-Task Learning; Convolutional Neural Network; Automatic modulation recognition; Feature Fusion

Article

IDAF: Iterative Dual-Scale Attentional Fusion Network for Automatic Modulation Recognition

Bohan Liu [1,*], Ruixing Ge [1], Yuxuan Zhu [1], Bolin Zhang [2], Xiaokai Zhang [3] and Yanfei Bao [1]

[1] Institute of Systems Engineering, Academy of Military Science of the People's Liberation Army, Beijing 100083, China
[2] National Key Laboratory of Science and Technology on Communication, University of Electronic Science and Technology of China, 611731, Chengdu, China
[3] College of Communications and Engineering, Army Engineering University of PLA, Nanjing 210007, China
* Correspondence: bohanliu@sina.com

Abstract: Recently, deep learning models have been widely applied to modulation recognition, which have become a hot topic due to their excellent end-to-end learning capabilities. However, current methods are mostly based on uni-modal inputs, which suffer from incomplete information and local optimization. To complement the advantages of different modalities, we focus on the multi-modal fusion method. Therefore, we introduce an iterative dual-scale attentional fusion (iDAF) method to integrate multimodal data. Firstly, two feature maps with different receptive field sizes are constructed using local and global embedding layers. Secondly, the feature inputs are iterated into the Iterative Dual Scale Attention Module (iDCAM), where the two branches capture the details of high-level features and the global weights of each modal channel, respectively. The iDAF not only extracts the recognition characteristics of each specific domains, but also complements the strengths of different modalities to obtain a fruitful view. Our iDAF achieves a recognition accuracy of 93.5% at 10dB and 0.6232 at full SNR. The comparative experiments and ablation studies effectively demonstrate the effectiveness and superiority of the iDAF.

Keywords: automatic modulation recognition; multimodal learning; convolutional neural network; attention mechanism

1. Introduction

Automatic modulation recognition [1,2] is the process of identifying the modulation of the received signal in the absence of sufficient a priori information. Defining the modulation is necessary for correct demodulation, which is fundamental in spectrum monitoring [3], information countermeasures [4], cognitive radio [5], etc. With the increasing development of wireless communication technology, the modulation of signals tends to be diversified, and the number of frequency-using devices is increasing. Therefore, the study of real-time and efficient AMR is of great practical significance.

The mainstream AMR methods are divided into two categories, i.e., likelihood theory-based (LB-AMR) [1,6,7] and the feature-based (FB-AMR) [2,8] methods. However, the performance of these traditional methods relies on manually estimated parameters [9], which leads to harder feature extraction under the high data transmission rates [10]. Instead of relying on artificial derivation to extract features, deep learning models feed signals directly into the network for end-to-end learning. Experiments have confirmed that the methods based on deep learning have better recognition accuracy than the traditional LB-AMR and FB-AMR methods [11]. At present, a large number of deep neural networks such as Convolutional Neural Network (CNN) [12], Denoising Automatic Encoder (DAE) [13], and Recurrent Neural Network (RNN) [14] are all introduced into AMR tasks. In the existing DL-AMR methods, most take a single modality as the input data type such as in-phase/quadrature (I/Q) [14], amplitude/phase series (A/P) [15], welch spectrum, square spectrum, and fourth power spectrum [16,17]. However, a single modality only contains the limited identifying information required for recognition completely from specific domains.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

For DL-AMR methods [12–14], different input data types have their own advantages. As shown in Table 1, input data from different modalities perform distinctively well for particular modulations due to the domain gap. Obviously, the I/Q, A/P, and spectral data have significant distinguishing abilities for PAM, QAM, and PSK modulations, respectively. However, the use of single-domain data formats does not provide the sufficiently efficient and complete view for recognition, which is due to the fact that different modes contain specific properties.

Table 1. Comparion of input data of different modalities.

Feature domains	Models	Effects
I/Q	CNN combined with Deep Neural Networks (DNN)[12], a combined CNN scheme[18]	achieve high recognition of PAM4 in low signal-to-noise ratio (SNR)
A/P	Long Short Term Memory (LSTM) [15], a LSTM denoising auto-encoder [13]	well recognize AM-SSB, and distinguish between QAM16 and QAM64[19]
Spectrum	RSBU-CW with welch spectrum, square spectrum, and fourth power spectrum [20], SCNN [17] with the short-time Fourier transform(STFT), a fine-tuned CNN modelčitezhang2019automatic with smooth pseudo-wigner-ville distribution and Born-Jordan distribution	achieve high accuracy of PSK [20], recognize OFDM well which is revealed only in the spectrum domain due to its plentiful sub-carriers [16]

In recent years, several studies have also focused on the advantages of multimodal information fusion for AMR tasks. In [21], modality discriminative features are captured separately using three Resnet networks, and I/Q, A/P, and amplitude of spectrum, square spectrum, and fourth power spectrum features are concatenated with the corresponding bitwise summation. [22] propose a dual-stream structure based on CNN-LSTM (DSCLDNN), which combines the characteristics of I/Q with A/P by pairwise cross-interacting the characteristics of the two streams. Specifically, the DSCLDNN multiplies I/Q and A/P features with an outer product. Unlike the above direct addition or multiplication fusion approach, [20] uses a PNN model to cross-fuse the three modal features in a fixed order. However, most of the above methods fuse multimodal features by direct or crosswise summation or outer product, which tends to ignore the variability of different modes and their different impacts on modulation identification.

Generally, the attention mechanism [23,24] can identify the channel-wise importance. Therefore, each modality has adaptively obtained its respective attention weight. For a feature map, attention weights need to be focused on both channel and spatial dimensions. Channel attention such as SENet [23], GSoPNet [25], and SRM [26] extract the attention information of different channels to distribute greater weight to important channels. For the spatial dimension, the attention mechanism such as GENet [24], RAM [27], and self-attention [28] are used to extract important spatial regions or spatial locations of high relevance. For multi-channel inputs composed of multimodal signals, the structure of the channel and spatial attention mechanisms was borrowed for the dual-channel attention fusion(DAF) we designed. Specifically, the dual channels are local and global branches. On the local branch, the spatial attention mechanism extracts local high-level feature details, while the channel attention mechanism on the global branch assigns attention weights to the different modal channels.

The main contribution of this work can be summarized as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

- We propose a deep learning method based on iterative dual-scale attentional fusion (iDAF), which complements the properties and complementarity of multimodal information with each other to achieve better recognition.
- We design two embedding layers to extract the local and global information, extracting information that promotes recognition from different-sized respective fields. The extracted features are sent into the iterative Dual-scale channel attention module (iDCAM), which consist of the local and global branch. The branches respectively focus on the details of the high-level features and the variability across modalities.
- Experiments on the RML2016.10A dataset demonstrate the validity and rationalization of iDAF. The highest accuracy amount of 93.5% is achieved at 10dB and the recognition accuracy is 0.6232 at full SNR.

2. Related works

2.1. Research on traditional AMR methods

For the AMR task, two traditional methods are LB-AMR and FB-AMR. The LB-AMR typically uses probability theory and hypothesis testing theory, while the FB-AMR is achieved by selecting representative features that best reflect the differences between modulated signals. LB-AMR mainly includes the average likelihood ratio test [2,29], the generalized likelihood ratio test [30], and the mixed likelihood ratio test [31]. However, although the likelihood technique is optimal in the sense of minimizing the probability of misclassification, the practical implementation is affected by computational complexity [15] and it is difficult to determine the appropriate analytic solution to the decision function [2]. In contrast, FB-AMR has low computational cost and achieves near-optimal performance, which has been proved the validity of extracted features through mathematical calculation [19,32,33]. However, the performance of the traditional algorithm relies on manually estimated parameters [9], and gets increasingly harder to extract features with the development of high data transmission rates [10]. Therefore, the strong automatic learning capability of DL models is widely used to accomplish the AMR task.

2.2. Study of different inputs and DL-models

DL-AMR methods achieve high recognition accuracy with different input modalities. By analyzing the I/Q vector, [12] based on a simple convolutional network (CNN) achieves higher accuracy in full SNR than traditional methods. In [10], the feedforward deep neural network (DNN) was used for pre-training, and the I and Q components were passed through an independent automatic encoder to realize unsupervised feature learning. OFDM signals were converted into I/Q samples in [9], and frequency domain analysis (FDA) pretreatment and $l2$ regularization were used to achieve high classification accuracy under low SNR. In addition to one-dimensional data, two-dimensional visual representations such as time-frequency and constellation images also show strong representation ability. A quarter spectrum diagram (Q-spectrum diagram) representation is proposed in [3], which is used by well-known convolutional neural networks such as VGG-16, AlexNet and ResNet18, respectively, with classification accuracy of more than 98% at high signal-to-noise ratio. [34] takes constellation map as the input of InceptionResnetV2-TA network when the signal-to-noise ratio is 4dB, the recognition rate on three typical signals is 3% higher than other algorithms. However, current models focus on information in a single domain of time, phase, and frequency, which results in underutilizing multimodal signal data. Therefore, we consider complementing the advantages of different modalities based on attentional mechanisms, to facilitate obtaining a complete view of the signal.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

3. The Proposed Method

In this section, we first preprocess the initial data to obtain three modalities representation. Then we introduce iterative dual-scale attentional fusion (iDAF), consisting of data embedding layers and iterative dual-channel attention module (iDCAM).

3.1. Data Preprocessing

This paper aims to identify modulation in a single-input single-output radio transmission system (SISO). The receiver transmits signal s through transmission channel h to obtain the baseband transmission signal.

$$s(i) = A(i)e^{j(\omega l+\varphi)}s(i) + n(i), i = 1, 2, 3...N, \tag{1}$$

where s is the complex baseband signal transmitted by the transmitter under some modulation scheme, ω is the frequency offset, φ is the phase offset, A is the communication channel gain, n is the Additive Gaussian White Noise (AWGN), i represents the i-th value received. The purpose of the automatic modulation recognition task is to transmit signals through the baseband of the receiver and determine the pattern of modulation recognition, which can be classified as a $P(y = C_K|s)$ estimation problem for identifying K types of radio modulations.

The key to the recognition task is to obtain the effective features of the signal, while the representational ability of the features extracted by a single modality is limited especially in the case of low SNR. In order to cover the amplitude, phase, and spectrum characteristics required identifying for modulation recognition, three modalities are selected to ensure that the required identifying information is included. I/Q and A/P contain instantaneous amplitude, phase and frequency information as modality one (IQ) and modality two (AP), respectively. The welch spectrum, square spectrum, and fourth power spectrum selected as the third modality (SA) represent the spectral characteristics of the signal in the frequency domain.

Therefore, prior to input into the neural network, the original signal symbol is transferred to the three modal representations in the following ways:

- In-phase/orthogonal (IQ): Generally, the receiver stores the signal in the modality of I/Q to facilitate mathematical operation and hardware design, which is expressed as follows:

$$\begin{aligned} V_{IQ} = \begin{pmatrix} I \\ Q \end{pmatrix} &= \begin{pmatrix} Re[s(1), s(2), ..., s(N)] \\ Im[s(1), s(2), ..., s(N)] \end{pmatrix} \\ &= \begin{pmatrix} Re[1], Re[2], ..., Re[n] \\ Im[1], Im[2], ..., Im[n] \end{pmatrix} \end{aligned} \tag{2}$$

  where I and Q represent the in-phase and quadrature components, Re and Im refer to the real and imaginary parts of the signal, respectively.
- Amplitude/phase(AP): Calculate the instantaneous amplitude and phase of the signal, expressed as:

$$V_{AP} = \begin{pmatrix} A \\ P \end{pmatrix} = \begin{pmatrix} Amplitude(n) = \sqrt{Re^2[n] + Im^2[n]} \\ Phrase(n) = arctan\frac{Im[n]}{Re[n]} \end{pmatrix} \tag{3}$$

  where the values of n are $0, 1, 2, ..., N-1$.
- Spectrum (SP): The spectrum expresses the change of frequency over time, which is an important discrimination of different modulations. the calculation of the spectrum is expressed as:

$$V_{SP} = \left| \sum_{i=0}^{N-1} s(i)^n e^{-j2\pi ki/N} \right|, k = 0, 1, 2, ..., N \tag{4}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

where n represents the n-th power of the spectrum, including 1, 2, 4 which are corresponding to the welch spectrum, square spectrum, and fourth power spectrum. Here, M1 and M2 represent signal waveform and frequency, and M3 refers to signal time-frequency characteristics. The feature vectors of the three modalities were normalized into (batchsize x 128).

In order to observe the specific performance of modulation on different modalities, we plot the data of IQ, AP, and SP modality by 11 modulations. It can be seen that several modulations will be well classified similarly in different modalities, while specific modulations will behave distinctively in a single modality. Therefore, we introduce an attentional fusion to integrate the above similar and distinct features,.

3.2. *Iterative dual-channel attention fusion (iDAF)*

For the iDAF, we design with two data embedding layers to construct the local and global feature maps, then send it into an iterative dual-channel attention module (iDCAM) for attention weight assignment as shown in Figure 1.

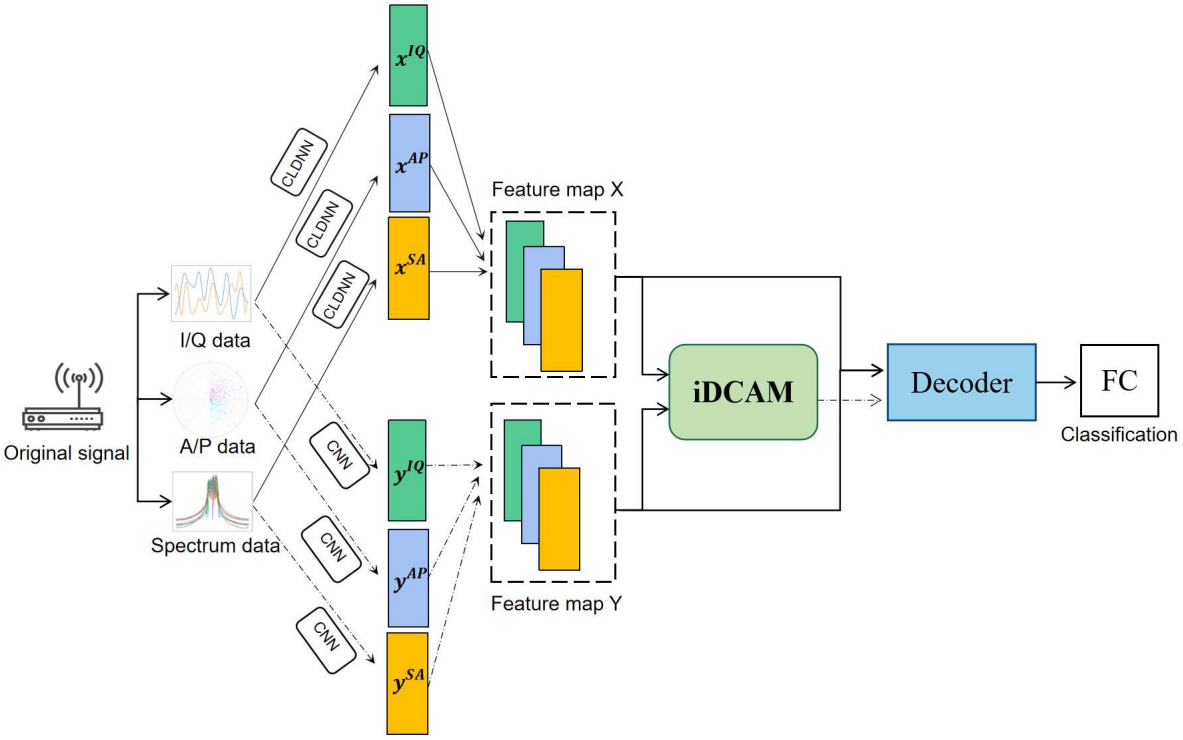


Figure 1. Architecture of the proposed iterative dual-scale attentional fusion (iDAF).

3.2.1. Data embedding

The signal data consists of three modal inputs, including I/Q, A/P, and spectrum analysis. For the original signal, it is preprocessed into three modalities inputs, denoted as $h_{IQ} \in \mathcal{R}^{128\times2}$, $h_{AP} \in \mathcal{R}^{128\times2}$, $h_{SP} \in \mathcal{R}^{128\times3}$. The preprocessed inputs h_m ($m \in (IQ, AP, SP)$) represent orthogonal information, amplitude-phase domain, and spectral features respectively.

Due to the variability of multimodal features, direct fusion would ignore the properties unique to different modalities. Therefore, we capture features from both local and global feature maps. The local feature map extracts detailed high-level semantic features, and the global feature map focuses on inter-modal salient characteristics. Therefore, we construct these two feature maps separately using feature extraction networks with different-sized receptive fields.

For the local feature map X, the feature extraction network is expected to focus on local details and contextual information. Inspired by [35], we propose the local embedding layer with CNN, LSTM

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

and DNN which is fine-tuned to extract local attention information. Firstly, preprocessed data passes through a few convolution layers to model frequency. Therefore, the long-term features are obtained by undistorted convolution (UD-Conv) layers with channel dimensions 128, 64, 32, and 16. Concretely, UD-Conv consists of a zero-padding layer of size (2,0,0,0), a convolution layer, the Relu function, and batch normalization. Using the zero-padding, two columns are added to ensure that the signal features can be transmitted with as little time-frequency information as possible. Following [36], the outputs of CNN are sent into LSTM and DNN. The LSTM layer is a bidirectional recursive model with 100 cells, which makes predictions using information both before and after the current moment in the sequence. The input is passed to the model in the original order, the incoming data in the reverse order, and finally the forward and reverse outputs are merged. The long-short time series learning capability of the LSTM identifies temporal correlations in I/Q data with inherent memory properties, and benefits learning the temporal dependencies of instantaneous amplitude and phase [15].

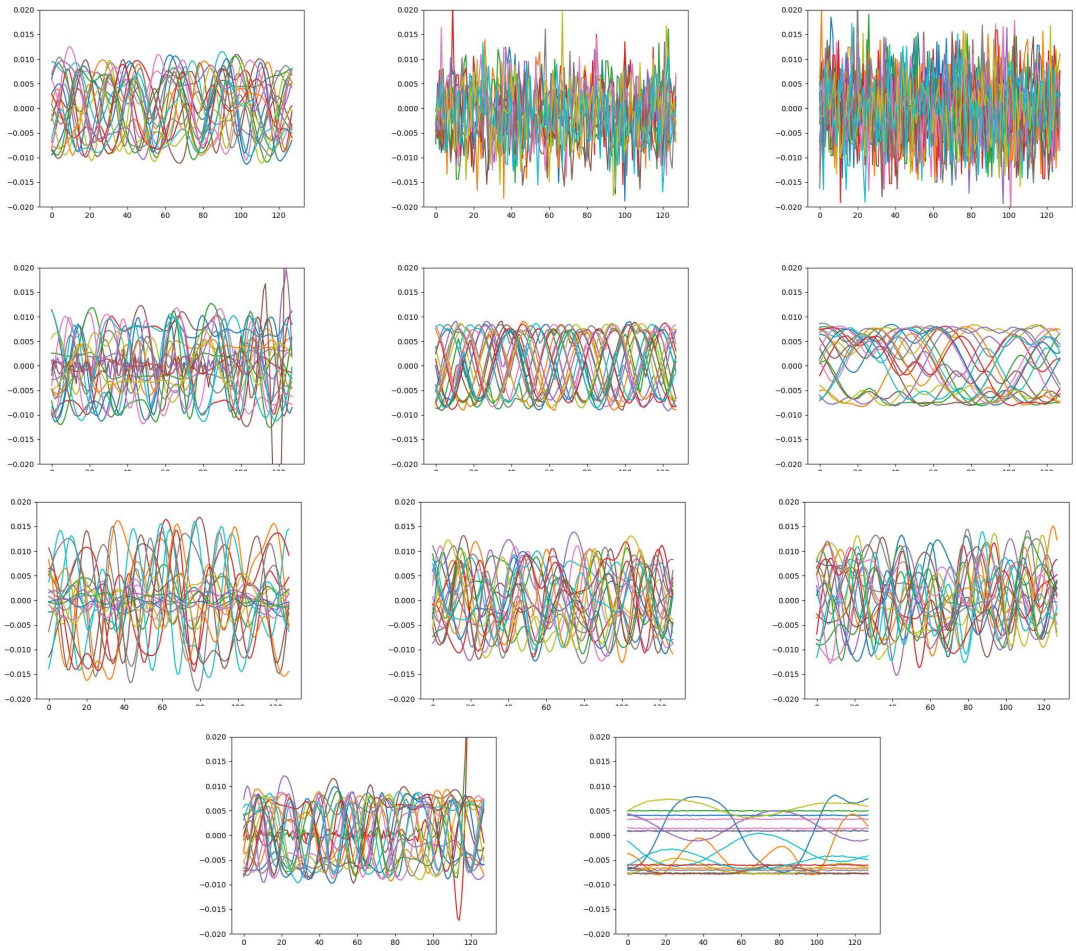


Figure 2. The IQ data plot of 11 modulations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

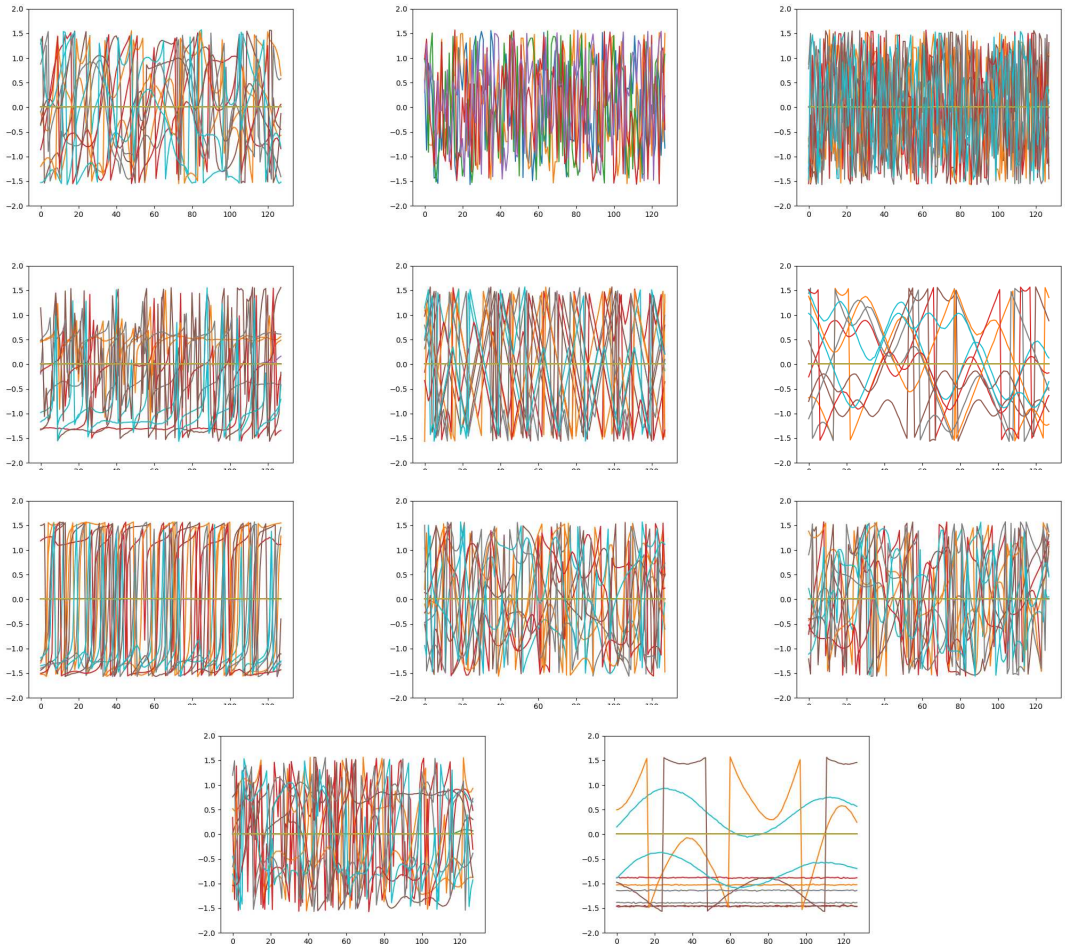


Figure 3. The AP data plot of 11 modulations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

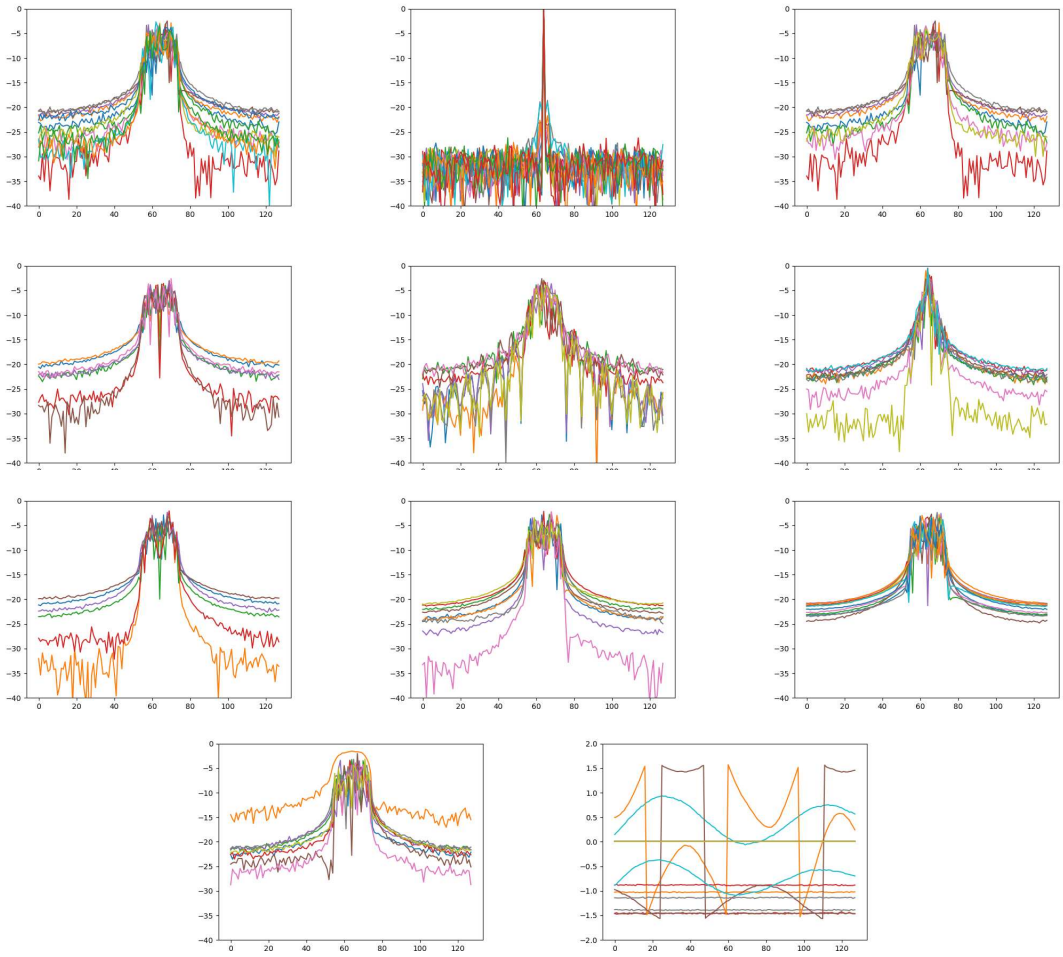
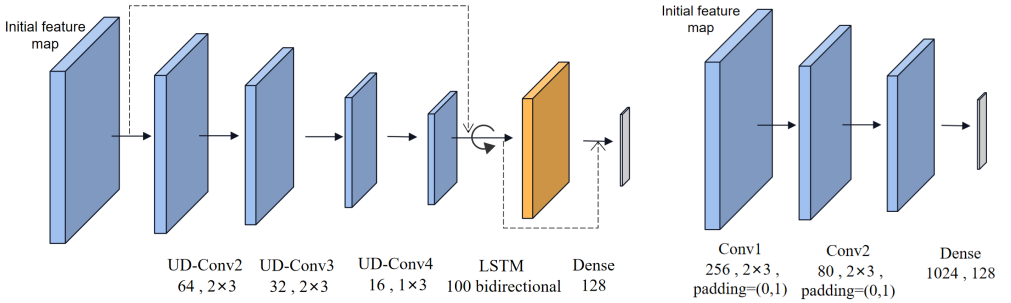


Figure 4. The SP data plot of 11 modulations.

The residual mapping function is a shortcut path between different layers, which can deepen the communication between deep and shallow neural network features. Inspired by [14], the Resnet has achieved the best performance on classifying signal modulation with a 4-convolution-layer structure. After four UD-Conv layers, long-term features are extracted by the convolution layers, while short-term information may be neglected during the convolution process. Therefore, the original data containing long-short-term features are entered into LSTM together with the extracted long-term features via the residual connection. Inspired by [37], the extracting capability of CNN is combined with LSTM and DNN. As shown in Figure 5(a), the learned short-term features are fed into the dense layer together with the long-term features previously extracted by CNN. The local embedding layer captures the data characteristics of each modal with unshared parameters, which is expressed as $x_m = E_x\left(h_m, \omega_m^{E_x}\right), (m \in (IQ, AP, SA))$, where E_x represents the local embedding layer and ω_m indicates the local network parameters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1



(a) The local embedding layer for feature map X.

(b) The global embedding layer for feature map Y.

Figure 5. The feature map construction networks.

To obtain the global feature map Y, an optimized CNN with 3 convolutional layers is utilized to extract features $y_m = E_y\left(h_m, \omega_m^{E_y}\right), (m \in (IQ, AP, SA))$ in the global receptive field in Figure 5(b).

3.2.2. Dual-scale channel attention module

After constructing the feature maps in the previous section, the feature maps are fed into an iterative multimodal attention module (iDCAM). The dual-scale channel attention module (DCAM) is a computational unit that can be constructed and superimposed for feature map transformation containing two branches.

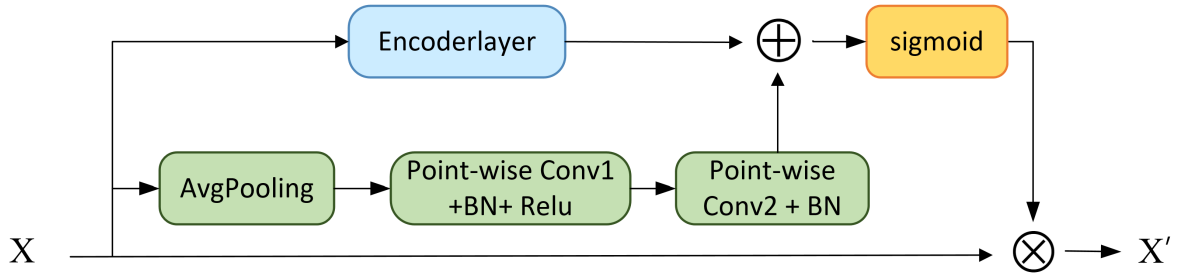


Figure 6. Architecture of the proposed Dual-scale channel attention module (DCAM).

The branches include a local attention branch and a global attention branch, correspondingly for extracting the local identification properties and the channel variability between modalities, respectively. The local attention branch extracts the intra-modal attention through the self-attention mechanism of the Transformer, which extracts the local recognition properties of specific modality features. Meanwhile, the global attention branch increases the receptive field by pooling to obtain inter-modal global attention in the channel dimension. The feature maps are respectively fed into the dual-scale channel attention module and the following steps are performed as follows:

1) Passing through the encoder.

To capture the attention information between different modalities, the feature map is first sent into the encoder layer of the Transformer [38]. The encoder consists of a self-attention module and a feed-forward neural network. Concretely, the self-attention mechanism is able to interact with the vectors converted from different sequence tokens, giving attention information about the correlation between different modalities. The basic formula of the self-attention mechanism is first expressed as follows:

$$\begin{cases} Q &= W_Q x \\ K &= W_K x \\ V &= W_V x \end{cases} \tag{5}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

Therefore, the input x is converted to a query Q, a key K, and a value V by means of three learnable weights W_Q,W_K, and W_V. Here, Q is used to query the similarity of other vectors to itself and K is used for indexing for operations.

By dot-multiplying Q and K, the similarity between the two is computed, which is then converted into a weight probability distribution to get the importance of different modalities in different signal sequences as attention information. Specifically, the attention information is normalized by scaling factor and softmax.

$$Atten(Q,K,V) = softmax(\frac{QK_T}{\sqrt{d_k}})V \tag{6}$$

Finally, the output of this self-attention layer is obtained by weighting the value V which helps in the classification with the attention information and then accumulating it. Utilizing multiple self-attention layer operations, the multi-head attention layer as shown in the following equation:

$$G(X) = Multihead(Q,K,V) = concat(head_1, head_2, head_n)W \tag{7}$$

$$\tag{8}$$

$$head_n = Atten(Q_n,K_n,V_n)s_nftmax(\frac{Q_nK_n^T}{\sqrt{d_k}})V_n \tag{9}$$

2) Construct the global channel attention matrix.

First, feature mappings across spatial dimensions H × W are aggregated after a squeeze compression operation. A channel descriptor containing global attention information is generated by global average pooling, which is denoted as follows:

$$z_i = F_{squ}(x_i) = \frac{1}{H \times W}\sum_{m=1}^{H}\sum_{n=1}^{H} x_i(m,n) \tag{10}$$

After squeeze compression, the aggregated information is sent into two convolution layers to capture the channel dependencies.

$$L(X) = F_{conv}(z) = B(Conv_1(\sigma(B(Conv_2(z))))) \tag{11}$$

where σ and B represent the Rectified Linear Unit (ReLU) function and Batch Normalization (BN), respectively. Specifically, the $Conv$ we used is the point-wise convolution, which enhances the nonlinear capabilities of the network. The kernel size of $Conv_1$ and $Conv_1$ is $1 \times 1 \times 1$ and $3 \times 1 \times 1$ respectively.

3) Matrix multiplication between the attention matrix and the original features.

$$H' = H \otimes W((X \oplus Y) = H \otimes \sigma(L(X) \oplus G(X)), \tag{12}$$

where $\otimes$ represents matrix addition and $\oplus$ is matrix multiplication. $W(X)$ contains the summation information of local attention $L(X)$ and global attention $G(X)$ extracted through DCAM.

3.2.3. Iteritive dual-scale attentional module

The inputs are high-level feature maps X and low-level feature maps Y. X utilizes the local sensing and context-sensitive inference capabilities of CNN and LSTM to capture the discriminative properties of each modality. However, the extracted high-level features are rich in local semantic information but ignore inter-modal difference information. In contrast, Y extracts global information with a larger perceptual field, and the extracted low-level features extract the distinctiveness between different modalities from a holistic perspective. However, due to the use of fewer convolutional layers,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

the deep feature semantic information is difficult to mine. Therefore, due to the desire to complement the advantages of low-level features and high-level features, an iterative dual-scale attentional module (iDCAM) is designed.

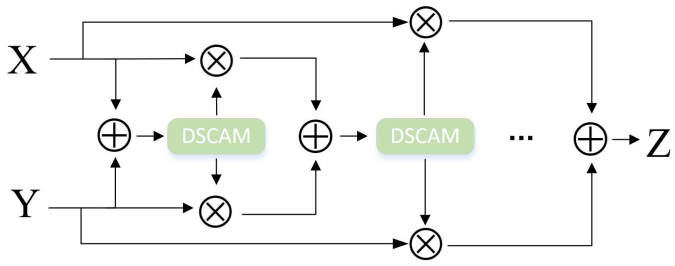


Figure 7. The iterative attention mechanism iDCAM.

By stacking the DCAM designed in the previous section, iDCAM assigns multimodal attention weights to different modality features.

$$Z_i = W_i(X \oplus Y) \otimes X + (1 - W_i(X \oplus Y) \otimes X, i = 1, 2, ...k \tag{13}$$

where $W(X \oplus Y)$ represents the summation information of local X and global Y.

3.2.4. Residual encoder

After passing through iDCAM, features are sent into the decoder along with the sum of intermediate features. The features are fed into the decoder after being assigned weights by iDCAM, and decoding is guided by a cross-attention mechanism using the sum of the intermediate features x_m ($m \in (IQ, AP, SA)$) and y_m ($m \in (IQ, AP, SA)$).

4. Experiment Results and Discussion

4.1. Datasets and implement details

The RML 2016.10A dataset contains eleven different types of modulation styles, including three analog and eight digital modulations. Specifically, the modulations are BPSK, QPSK, 8PSK, 16QAM, 64QAM, BFSK, CPFSK, PAM4, WB-FM, AM-SSB and AM-DSB, with SNRs of 20 values ranging from -20 dB to 18 dB. The total number of samples is 220,000 and each sample is a complex-valued with independent in-phase and quadrature (I/Q) parts. The data structure is of shape [2,128], where 2 corresponds to the I and Q channels and 128 corresponds to 128 sampling points. The channel is additive white Gaussian noise (AWGN) channel. In our experiments, the dataset is divided into the training set, testing set, and validation set in the ratio of 6:2:2.

The training and prediction of our iDAF model and other mainstream models are replicated on an Nvidia Tesla V100 in Pytorch deep learning algorithm platform. The optimizer used is Adam. The learning rate is set to 0.0001 and decreases in an orderly manner following an exponential decay. The precision score is used as the evaluation metric to evaluate the recognition capability of multiple modulated signals, which represents the ratio of the number of signals whose modulation is correctly recognized to the total number of signals.

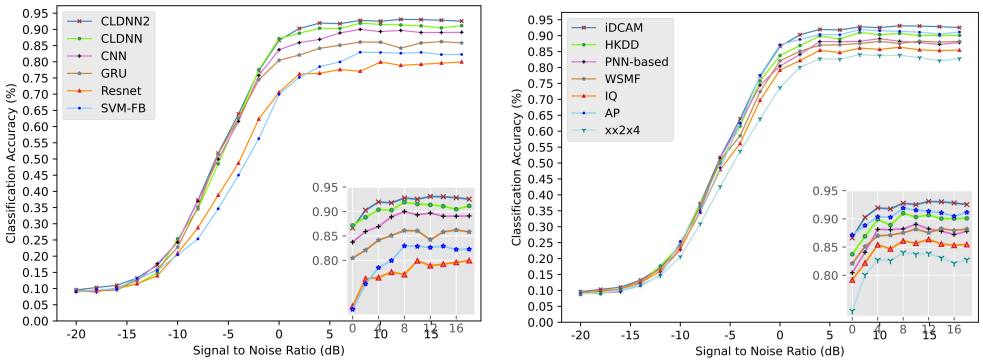
4.2. Comparative validity experiments

4.2.1. Compare local embedding layer with other feature extraction networks

To validate the effect of the local embedding layer, we conducted experiments to compare the recognition capabilities at several classical networks that extract features from signal data. Figure 8(a) illustrates the increasing trend of recognition accuracy with SNR from -20 to 18dB for different

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

models. As can be seen from the figure, classical FB-AMR models such as SVM-FB, Resnet, and ResCNN are unable to accurately extract signal features due to the lack of adaptation to the signal data. Without residual concatenation of the original data and the long-term features extracted by the CNN, the resulting feature map is not a holistic view that fuses multimodal long-term and short-term information. The highest accuracy amount of 93.5% is achieved at 10dB for iDCAM and the recognition accuracy is 0.6232 at full SNR.



(a) Comparisons of feature extraction models with local embedding layer.

(b) Comparision with uni-modal and Multimodal methods.

Figure 8. Comparision experiments.

4.2.2. Compare iDCAM and other attention mechanisms

In order to assign different weights to the features of different modalities, the attention layer not only needs to extract local high-level semantic information but also pay attention to the global attention of the channel. For the current attention methods, channel attention mechanism SENet [23], combined spatial attention mechanism BAM [39] and CBAM [40], multi-branch attentional network SKNet [41], and self-attention mechanism Transformer are compared to verify the superiority of the proposed iDCAM. For signal samples with an initial data shape of [2,128], information in the temporal dimension is not sufficient for recognition, but combining channel attention can effectively extract features. As shown in Table 2, SENet clearly outperforms SKNet when used alone, and CBAM with fused channel and spatial attention has better recognition accuracy. As for parameters that calculation required, iDCAM requires only a small number of parameters for accurate recognition, without relying on a backbone feature extraction network such as Resnet,.

Table 2. Comparison of multiple attention mechanisms.

Model	Accuracy	Params(M)
SENet-ResNet18	0.6032	11.9
SKNet-50	0.5994	27.6
CBAM-ResNeXt50	0.6082	27.8
Self-attention	0.618	63.5
BAM-Resnet-50	0.6038	24.7
iDCAM	0.6232	6.9

4.2.3. Comparison with unimodal and multimodal methods

In order to verify the effectiveness of the proposed multimodal approach, the effects of the IQ, AP, and SP unimodal inputs and multiple multimodal approaches are compared under the same experimental setup, respectively.

Firstly, from the comparison of single-modal confusion matrices in Figure 9(a), Figure 9(b), and Figure 9(c), it can be seen that the model effect of single-modal input without iDCAM performs poorly

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

on specific modulations. For example, the accuracy of IQ and SP on AM-SSB is less than 50%, and AP is prone to misclassify QAM64 as AM-SSB. After fusing multimodal features through our attention mechanism iDCAM in Figure 9(d), the strengths of different modalities compensate for each other and greatly improve the shortcomings.

(a) CLDNN [14] model with IQ input

(b) LSTM [15] model with AP input

(c) CNN model with spectrum input

(d) iDCAM model with multiple inputs

Figure 9. The recognition performance of different modes to 11 modulations.

Secondly, the comparison with other multimodal methods also validates the effectiveness of iDCAM. As shown in Figure 8(b), iDCAM leads most of the methods, and the recognition accuracy at full signal-to-noise ratio exceeds that of the HKDD method by 1.5%. We attribute this to iDCAM's ability to adaptively assign local-global attention weights to multimodal features, as opposed to other methods that directly connect or cross-connect features in a fixed order.

4.3. Abaltion studys

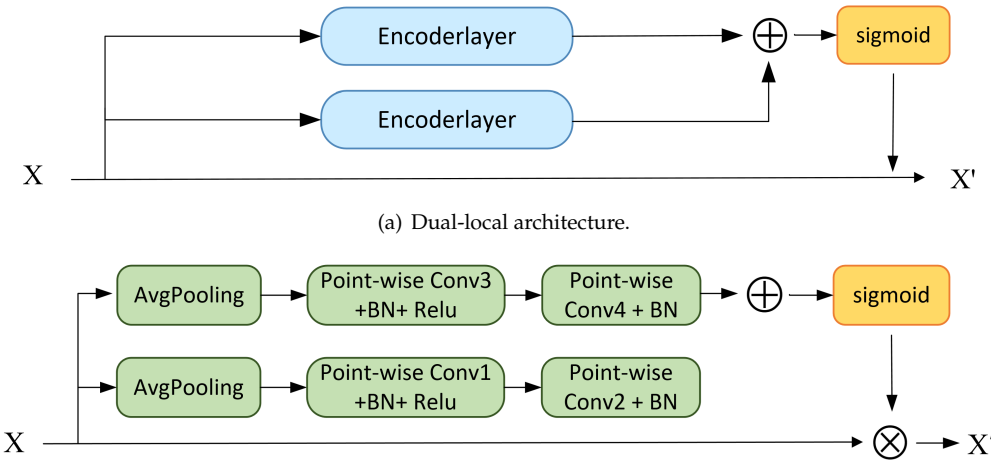
4.3.1. Ablation experiments at different scales with DCAM

First, single-scale comparisons are performed to verify the superiority of the dual scale. For our proposed dual-scale approach, local scale and global scale extract attentional information from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2023 doi:10.20944/preprints202309.0234.v1

different receptive fields. As shown in Table 2, a single scale leads to the loss of local details or global attention when designing the attention extractor with only local or global branches. The local branch is about 1% more accurate than the global branch. This indicates that the local structure focuses on high-level features and ignores the specificity of different modality data, while the global structure fails to achieve accurate recognition due to the lack of feature details.

Second, ablation experiments were performed at different dual scales. Specifically, we replicate the individual local or global attention mechanisms separately to compose a dual-local and dual-global structure. As shown in the Figure 10, dural-local and dural-global represent that both branches are set to the same receptive field. From the results of the ablation experiments in Table 3, it seems that the dual-branch structure is generally better than the single-branch ones. Model recognition works best only when both local and global information are fused attentively, while the amount of calculations is relatively reasonable. The two-branch structure increases the FLOPs compared to the single branch, but maintains a controlled growth rather than an unacceptable one.



(a) Dual-local architecture.

(b) Dual-global architecture.

Figure 10. Comparision with different dual structure.

Table 3. The ablation results of different scales.

Architectures	Recognition accuracy	FLOPs (G)
Local	0.618	10.1
Global	0.6081	/
Dual-local	0.6192	20.2
Dual-global	0.6104	/
Local-global	0.6232	10.9

4.3.2. Ablation experiments with iterative layers of iDSACM

In order to obtain the number of iteration layers for optimal recognition results of iDCAM, one, two, three and four iterations of iDCAM are used for modulation recognition respectively. As shown in Table 4, the best results can be achieved with two iterations. When DCAM is not iterated (i.e., the number of layers is one), attentional information extraction is not adequately extracted. In contrast, the deep network leads to a decrease in the correctness rate when the number of layers is too high.

Table 4. The ablation results of iterative layers.

Iterations K	one-layer	two-layer	three-layer	four-layer
Accuracy	0.6194	0.6232	0.6204	0.6181

5. Conclusion

In this paper, we introduce an iterative dual-scale attentional fusion (iDAF) method to integrate multimodal data. In the proposed method, we realize significant classification superior to the other fusion DL-AMR methods, and achieve a recognition accuracy of 93.5% at 10dB and 0.6232 at full SNR. In future work, one promising direction is to further mine the deeper characteristics of different modalities, and demonstrate the reason for the existence of variability in different modalities by means of mathematical analysis.

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

References

1. Dai, A.; Zhang, H.; Sun, H. Automatic modulation classification using stacked sparse auto-encoders. In Proceedings of the 2016 IEEE 13th international conference on signal processing (ICSP). IEEE, 2016, pp. 248–252.
2. Al-Nuaimi, D.H.; Hashim, I.A.; Zainal Abidin, I.S.; Salman, L.B.; Mat Isa, N.A. Performance of feature-based techniques for automatic digital modulation recognition and classification—A review. *Electronics* **2019**, *8*, 1407.
3. Bhatti, F.A.; Khan, M.J.; Selim, A.; Paisana, F. Shared spectrum monitoring using deep learning. *IEEE Transactions on Cognitive Communications and Networking* **2021**, *7*, 1171–1185.
4. Richard, G.; Wiley, E. The interception and analysis of radar signals. *Artech House, Boston* **2006**.
5. Kim, K.; Spooner, C.M.; Akbar, I.; Reed, J.H. Specific emitter identification for cognitive radio with application to IEEE 802.11. In Proceedings of the IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference. IEEE, 2008, pp. 1–5.
6. Wei, W.; Mendel, J.M. Maximum-likelihood classification for digital amplitude-phase modulations. *IEEE transactions on Communications* **2000**, *48*, 189–193.
7. Xu, J.L.; Su, W.; Zhou, M. Likelihood-ratio approaches to automatic modulation classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **2010**, *41*, 455–469.
8. Hazza, A.; Shoaib, M.; Alshebeili, S.A.; Fahad, A. An overview of feature-based methods for digital modulation classification. In Proceedings of the 2013 1st international conference on communications, signal processing, and their applications (ICCSPA). IEEE, 2013, pp. 1–6.
9. Hao, Y.; Wang, X.; Lan, X. Frequency Domain Analysis and Convolutional Neural Network Based Modulation Signal Classification Method in OFDM System. In Proceedings of the 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP). IEEE, 2021, pp. 1–5.
10. Ali, A.; Yangyu, F. Unsupervised feature learning and automatic modulation classification using deep learning model. *Physical Communication* **2017**, *25*, 75–84.
11. Chang, S.; Huang, S.; Zhang, R.; Feng, Z.; Liu, L. Multitask-learning-based deep neural network for automatic modulation classification. *IEEE internet of things journal* **2021**, *9*, 2192–2206.
12. 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, September 2-5, 2016, Proceedings 17. Springer, 2016, pp. 213–226.
13. Ke, Z.; Vikalo, H. Real-time radio technology and modulation classification via an LSTM auto-encoder. *IEEE Transactions on Wireless Communications* **2021**, *21*, 370–382.
14. Liu, X.; Yang, D.; El Gamal, A. Deep neural network architectures for modulation classification. In Proceedings of the 2017 51st Asilomar Conference on Signals, Systems, and Computers. IEEE, 2017, pp. 915–919.
15. Rajendran, S.; Meert, W.; Giustiniano, D.; Lenders, V.; Pollin, S. Deep learning models for wireless signal classification with distributed low-cost spectrum sensors. *IEEE Transactions on Cognitive Communications and Networking* **2018**, *4*, 433–445.
16. Zhang, Z.; Wang, C.; Gan, C.; Sun, S.; Wang, M. Automatic modulation classification using convolutional neural network with features fusion of SPWVD and BJD. *IEEE Transactions on Signal and Information Processing over Networks* **2019**, *5*, 469–478.

17. Zeng, Y.; Zhang, M.; Han, F.; Gong, Y.; Zhang, J. Spectrum analysis and convolutional neural network for automatic modulation recognition. *IEEE Wireless Communications Letters* **2019**, *8*, 929–932.
18. Shi, F.; Hu, Z.; Yue, C.; Shen, Z. Combining neural networks for modulation recognition. *Digital Signal Processing* **2022**, *120*, 103264.
19. Fu-qing, H.; Zhi-ming, Z.; Yi-tao, X.; Guo-chun, R. Modulation recognition of symbol shaped digital signals. In Proceedings of the 2008 International Conference on Communications, Circuits and Systems. IEEE, 2008, pp. 328–332.
20. Zhang, X.; Li, T.; Gong, P.; Liu, R.; Zha, X. Modulation recognition of communication signals based on multimodal feature fusion. *Sensors* **2022**, *22*, 6539.
21. Qi, P.; Zhou, X.; Zheng, S.; Li, Z. Automatic modulation classification based on deep residual networks with multimodal information. *IEEE Transactions on Cognitive Communications and Networking* **2020**, *7*, 21–33.
22. Zhang, Z.; Luo, H.; Wang, C.; Gan, C.; Xiang, Y. Automatic modulation classification using CNN-LSTM based dual-stream structure. *IEEE Transactions on Vehicular Technology* **2020**, *69*, 13521–13531.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
24. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. *Advances in neural information processing systems* **2018**, *31*.
25. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2019, pp. 3024–3033.
26. Lee, H.; Kim, H.E.; Nam, H. Srm: A style-based recalibration module for convolutional neural networks. In Proceedings of the Proceedings of the IEEE/CVF International conference on computer vision, 2019, pp. 1854–1862.
27. Mnih, V.; Heess, N.; Graves, A.; et al. Recurrent models of visual attention. *Advances in neural information processing systems* **2014**, *27*.
28. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
29. Yang, F.; Yang, L.; Wang, D.; Qi, P.; Wang, H. Method of modulation recognition based on combination algorithm of K-means clustering and grading training SVM. *China communications* **2018**, *15*, 55–63.
30. Hussain, A.; Sohail, M.; Alam, S.; Ghauri, S.A.; Qureshi, I.M. Classification of M-QAM and M-PSK signals using genetic programming (GP). *Neural Computing and Applications* **2019**, *31*, 6141–6149.
31. Das, D.; Bora, P.K.; Bhattacharjee, R. Blind modulation recognition of the lower order PSK signals under the MIMO keyhole channel. *IEEE Communications Letters* **2018**, *22*, 1834–1837.
32. Liu, Y.; Liang, G.; Xu, X.; Li, X. The Methods of Recognition for Common Used M-ary Digital Modulations. In Proceedings of the 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008, pp. 1–4. https://doi.org/10.1109/WiCom.2008.410.
33. Benedetto, F.; Tedeschi, A.; Giunta, G. Automatic Blind Modulation Recognition of Analog and Digital Signals in Cognitive Radios. In Proceedings of the 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall), 2016, pp. 1–5. https://doi.org/10.1109/VTCFall.2016.7880915.
34. Jiang, K.; Zhang, J.; Wu, H.; Wang, A.; Iwahori, Y. A novel digital modulation recognition algorithm based on deep convolutional neural network. *Applied Sciences* **2020**, *10*, 1166.
35. Sainath, T.N.; Vinyals, O.; Senior, A.; Sak, H. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4580–4584. https://doi.org/10.1109/ICASSP.2015.7178838.
36. Sermanet, P.; LeCun, Y. Traffic sign recognition with multi-scale convolutional networks. In Proceedings of the The 2011 international joint conference on neural networks. IEEE, 2011, pp. 2809–2813.
37. Soltau, H.; Saon, G.; Sainath, T.N. Joint training of convolutional and non-convolutional neural networks. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 5572–5576.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
39. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514* **2018**.

40. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
41. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510–519.