Learning Cross-Domain Features With Dual-Path Signal Transformer

Lei Zhai, Yitong Li, Zhixi Feng, Shuyuan Yang, *Senior Member, IEEE*, and Hao Tan

***Abstract*— The past decade has witnessed the rapid development of deep neural networks (DNNs) for automatic modulation classification (AMC). However, most of the available works learn signal features from only a single domain via DNNs, which is not reliable enough to work in uncertain and complex electromagnetic environments. In this brief, a new cross-domain signal transformer (CDSiT) is proposed for AMC, to explore the latent association between different domains of signals. By constructing a signal fusion bottleneck (SFB), CDSiT can implicitly fuse and classify signal features with complementary structures in different domains. Extensive experiments are performed on RadioML2016.10A and RadioML2018.01A, and the results show that CDSiT outperforms its counterparts, particularly for some modulation modes that are difficult to classify before. Through ablation experiences, we also verify the effectiveness of each module in CDSiT.**

***Index Terms*— Automatic modulation classification (AMC), cross-domain transformer, multimodal learning, signal transformer (SiT).**

I. Introduction

Automatic modulation classification (AMC) is critical in noncooperative communication systems such as radio spectrum resources monitoring and modern electronic warfare [1]. Traditional AMC methods can generally be divided into two categories: likelihood-based [2], [3] and feature-based [4] methods. The estimated results of likelihood-based methods rely heavily on a large number of observations, which will result in high computational complexity. The feature-based methods need tedious "feature engineering" and the dense modulation schemes in modern communication systems often lead to unsatisfactory performance [5], [6], [7], [8], [9].

In very recent years, deep neural networks (DNNs) have been developed for simultaneous feature extraction and modulation classification in an "end-to-end" manner. Compared to traditional methods, they do not need any signal prior and have rapid predication. Various types of DNNs, including convolutional neural networks (CNNs) [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], recurrent neural networks (RNNs) [20], hybrid networks [21], [22], and Transformers [23], have been developed for AMC. Although these DNN-based AMC methods achieve significant performance gains [24], [25], [26], [27], [28], [29], [30], most of them deal with information from a single domain. For example, TRNN [23] divides the in-phase and quadrature (I/Q) signal of length n into several short sequences of equal length whose size is $(2a, b)$, $a \times b = n$ as the inputs. The I/Q components of signal waveforms are adopted for modulation classification in [10], [11], [12], and [21]. The time–frequency spectrum is fed into LSTM2 for classification [20], and the cyclic spectrograms of signals are employed in [18]. ICRNs [14] adopts the Choi–Williams distribution image of signals. Other descriptions of signals are also considered, such as the constellation diagrams [13], [15], high-order cumulants [17], and multitiming constellation diagrams [16].

It is well known that modulation signals can be characterized from various aspects, including amplitude, phase, frequency, power spectral density, and constellation diagram. Each of these aspects reflects a unique view of modulation type. However, one single view could not reliably and stably reflect the intrinsic modulation characteristics in uncertain and complex electromagnetic environments. For instance, the waveforms of modulation signals will exhibit dramatic changes as the modulation parameters vary. Also, the constellation shape and direction of modulation signals, together with the signal strength and spectral width, will change significantly in the presence of multipath effects. On the other hand, signals with similar modulation types, such as QAM32 and QAM64 signals, may look very similar in their constellation diagrams and waveforms. Therefore, in order to accurately identify the modulation types in complex environments, it is necessary to integrate the signal descriptions from multiple domains.

Some pioneered works have been advanced for AMC by cascading multidimensional descriptions of signals. For example, WSFM [31] combines multiple signal modalities by concatenating intermediate feature maps in CNN. In Ms-RaT [32], the amplitude and phase patches are extracted from the signal spectrum and then concatenated and fed into a multiscale network. However, they all adopt simple feature slicing for multidomain fusion, which could not well explore the association among multiple modalities.

As a recent research hotspot in machine learning, multimodal deep learning (MDL) involves relating information and learning features from multiple modalities via deep networks [24], [25], [26], [27], [28], [29], [30], [33], [34]. It has been theoretically proven that MDL is more powerful than single-modal deep learning since it can learn better embedding via interaction among modalities [35]. Thus, multidomain description of modulation signals can be combined together for learning more discriminative features. Inspired by it, in this brief, a new cross-domain signal transformer (CDSiT) is constructed, where a dual-path structure is used to learn cross-domain features of signals for AMC. CDSiT primarily consists of signal embedding (SE), signal transformer (SiT), and signal fusion bottleneck (SFB). These modules are carefully designed to explore the internal association between diverse characterization of signals. Extensive experiments are carried out on several benchmark datasets to verify the effectiveness of CDSiT.

Different from the available DNN-based AMC methods, the main contributions of our work are summarized as follows.

1) MDL is introduced into AMC, to develop a new dual-path SiT for cross-domain feature learning and classification of signals. To the best of our knowledge, this is the first MDL-based AMC work with transformer-like architecture.
2) CDSiT model is carefully designed for fine-grained representation and cross-domain fusion of multimodal signal sequences, in which multihead self-attention and convolution are used complementarily to capture long-range dependence and details in signals.

Manuscript received 30 November 2022; revised 26 May 2023 and 18 November 2023; accepted 30 December 2023. Date of publication 23 January 2024; date of current version 6 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62171357, Grant U22B2018, Grant 62276205, and Grant 61906145; and in part by the Foundation of Intelligent Decision and Cognitive Innovation Center of State Administration of Science, Technology and Industry for National Defense, China. *(Corresponding authors: Zhixi Feng; Shuyuan Yang.)*

The authors are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: raezhailei@gmail.com; ytlee@stu.xidian.edu.cn; zhxfeng2013@gmail.com; syyang@xidian.edu.cn; htan_1@stu.xidian.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2024.3350609

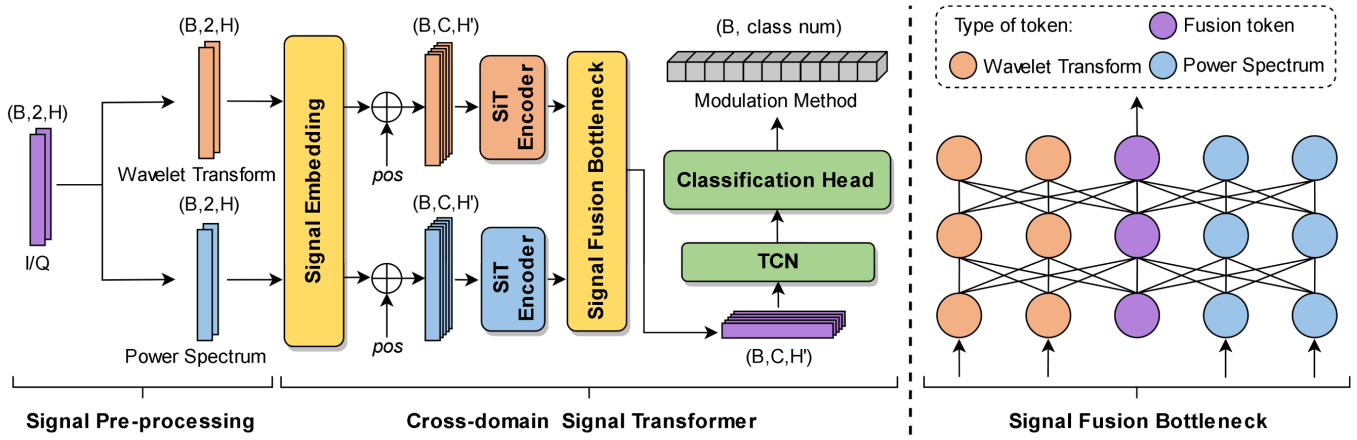

Fig. 1. Structure of CDSiT for fusing wavelet transform results of signals and their power spectrum.

3) Besides the combination mentioned in this brief, the proposed method is extensible and can work with a wide range of modalities. It is validated on some benchmark datasets.

II. Dual-Path SiT

A. Problem Formulation

Assume that the received baseband signal $\{x_1, x_2, \ldots, x_N\}$ is transmitted over an additive white Gaussian noise (AWGN) channel and sampled from pulses that meet the Nyquist limit, which can be written as

$$x_i(h) = \alpha_h e^{j(\Delta\phi + \pi \Delta f h/H)} S_{c_i} + w(h), \quad h = 1, 2, \ldots, H \tag{1}$$

where α_h is the time-varying amplitude gain of the channel based on the Rayleigh distribution in the limit of $(0, 1]$. $\Delta\phi$ represents the carrier phase offset that follows the uniform distribution $U(0, \pi/16]$. Δf is the normalized carrier frequency offset. $S_{c_i}(h)$, $c_i = 1, 2, \ldots, C$, stands for the hth transmitted symbol extracted from the constellations of the c_ith modulation type, and C is the total number of modulation types. w is the AWGN with mean 0 and variance σ_w^2. H is the length of the signal.

For the received baseband signal, our goal is to make the predicted distribution of the modulation scheme approach its true distribution, which can be described as

$$\arg\min_{\theta} -\frac{1}{N} \sum_{i=1}^{N_c} \sum_{c=1}^{C} y_i^{(c)} \log f_\theta(x_i) \tag{2}$$

where $N = \sum_{c=1}^{C} N_c$, in which N_c is the number of observed signals for each modulation type, $f_\theta(\cdot)$ represents a classifier with learnable parameters θ, and $y_i^{(c)}$ denotes the modulation type label of the ith received signal with the cth modulation type.

B. Architecture of SiT

It is well known that modulation signals are typically long-time sequences, which is suitable for transformer to model long-range temporal dependence. Some works have shown that if a network uses several convolution layers while maintaining a pure transformer structure, its dependence on large amounts of data can be reduced [36]. Thus, in our work, we design a new SiT primarily based on multihead self-attention and convolutions over signals. First, an SE module consisting of multiple 1-D convolution layers is constructed, to extend the dimensionality of inputs and mine the interaction information between the input channels. The convolutional layer allows the model to retain local spatial information, and SiT can achieve fine-grained representation of signal sequences for better SE. Furthermore, this model need not map the signal to a specific dimensionality and can deal with signals with arbitrary length, so it has more flexibility than other models, such as ViT [25], [26], [27], [28], [29], [30]. Then, temporal convolution is adopted to further explore the temporal characteristics of signals.

Given a signal $x \in \mathbb{R}^{2\times H}$, we first formulate the signal token $\mathbf{z}^0 \in \mathbb{R}^{C'\times H'}$ by

$$\mathbf{z}^0 = f^{\mathbf{E}}(x) + \eta_0^{\text{pos}} \tag{3}$$

where SE $f^{\mathbf{E}}(\cdot)$ is implemented via convolutional layers, which can extract the interaction features among channels. C' is the number of channels, and η_0^{pos} is a position encoding for preserving position information [37] and can be physically encoded as

$$\eta_0^{\text{pos}}(i, j) = \begin{cases} \sin\left(i/10000^{2j/H'}\right) \\ \cos\left(i/10000^{(2j-1)/H'}\right) \end{cases} \tag{4}$$

where $j \in \{1, 2, \ldots, H'/2\}$, $i = \text{pos}_i \in \{0, 1, \ldots, C'\}$, and (i, j) is the jth position of the ith component.

Then, we feed the signal token $\mathbf{z}^0$ into the constructed SiT encoder $f^{\text{SiT}}(\cdot)$, which consists of multiheaded self-attention (MSA), multilayer perceptron (MLP), and layer normalization (LN). At the lth layer, $\mathbf{z}^l = f^{\text{SiT}}(\mathbf{z}^{l-1})$ can be expressed as

$$\begin{aligned} \tilde{\mathbf{z}}^l &= \text{MSA}\big(\text{LN}\big(\mathbf{z}^{l-1}\big)\big) + \mathbf{z}^{l-1}, \quad l = 1, \ldots, L \\ \mathbf{z}^l &= \text{MLP}\big(\text{LN}\big(\tilde{\mathbf{z}}^l\big)\big) + \tilde{\mathbf{z}}^l, \quad l = 1, \ldots, L. \end{aligned} \tag{5}$$

Finally, the prediction $y \in \mathbb{R}^C$ can be obtained through the classification head with a temporal convolution network (TCN) block $f^{\text{TCN}}(\cdot)$, which can be expressed as

$$y = \text{LN}(f^{\text{TCN}}(\mathbf{z}^L)). \tag{6}$$

C. Signal Fusion Bottleneck

Cross-modal attention is commonly utilized in multimodal fusion, where the query, value, and key are derived from various modalities. In this brief, cross-modal attention is utilized to design a SFB f^{SFB} with a set of randomly initialized fusion tokens $\mathbf{z}_{\text{fst}}^0$ to realize the interaction among multidomain information. For example, fusing the wavelet transform and the power spectrum of the signal, the SBF f^{SBF} can be calculated as

$$\begin{aligned} \left[\mathbf{z}_{\text{pow}}^{l+1} \| \hat{\mathbf{z}}_{\text{fst}}^{l+1}\right] &= f_{\theta_{\text{pow}}}^{\text{SiT}}\left(\left[\mathbf{z}_{\text{pow}}^{l} \| \mathbf{z}_{\text{fst}}^{l}\right]\right) \\ \left[\mathbf{z}_{\text{wav}}^{l+1} \| \mathbf{z}_{\text{fst}}^{l+1}\right] &= f_{\theta_{\text{wav}}}^{\text{SiT}}\left(\left[\mathbf{z}_{\text{wav}}^{l} \| \hat{\mathbf{z}}_{\text{fst}}^{l+1}\right]\right) \end{aligned} \tag{7}$$

Algorithm 1 Forward Propagation of CDSiT

Input: Wavelet transform x_{wav}, Power spectrum x_{pow}, Layer number of SiT encoder L_f, Layer number of CDSiT L.
Output: The signal label vector y

Random initialization the fusion token $\mathbf{z}_{\text{fst}}^0$;
Calculate tokens $\mathbf{z}_{\text{wav}}^0$ and $\mathbf{z}_{\text{pow}}^0$ from x_{wav} and x_{pow} as in eq. (3), respectively;
for $l = 1, 2, \ldots, L$ **do**
  if $l \leq L_f$ **then**
    $\mathbf{z}_{\text{pow}}^{l+1} = f^{\text{SiT}}(\mathbf{z}_{\text{pow}}^l)$ as in eq. (5);
    $\mathbf{z}_{\text{wav}}^{l+1} = f^{\text{SiT}}(\mathbf{z}_{\text{wav}}^l)$ as in eq. (5);
  else
    $[\mathbf{z}_{\text{pow}}^{l+1}, \mathbf{z}_{\text{wav}}^{l+1}, \mathbf{z}_{\text{fst}}^{l-L_f}] = f^{\text{SFB}}\left(\mathbf{z}_{\text{pow}}^l, \mathbf{z}_{\text{wav}}^l, \mathbf{z}_{\text{fst}}^{l-L_f-1}; \theta_{\text{pow}}, \theta_{\text{wav}}\right)$ as in eq. (7);
$y = \text{LN}(f^{\text{TCN}}(\mathbf{z}^L))$ as in eq. (6);
return y.

where $\hat{\mathbf{z}}_{\text{fst}}^{l+1}$ is an intermediate variable of $\mathbf{z}_{\text{fst}}^{l+1}$, $\mathbf{z}_{\text{pow}}^l$ is the input power spectrum sequence of the lth SiT layers $f_{\theta_{\text{wav}}}^{\text{SiT}}$, and $\mathbf{z}_{\text{wav}}^l$ is the input wavelet transform input sequence of the lth SiT layers $f_{\theta_{\text{pow}}}^{\text{SiT}}$.

D. Cross-Domain SiT

As shown in Fig. 1, the proposed CDSiT mainly consists of SE, SiT encoder, SFB, and TCN block. The wavelet transform and the power spectrum of signals are first fed into the SE module. Then, the first L_f layer of CDSiT is made up of two SiT encoders, $f_{\theta_{\text{wav}}}^{\text{SiT}}$ and $f_{\theta_{\text{pow}}}^{\text{SiT}}$. Finally, the fusion and classification of the two domain features are then completed via SFB, TCN, and a linear classification head. A detailed description of the forward propagation of CDSiT is shown in Algorithm 1.

III. Experiment Results

A. Dataset

In this section, we investigate the performance of the proposed CDSiT on the RadioML2016.10A dataset [40] and the RadioML2018.01A dataset [10]. They are generated with GNU radio project, with local oscillator (LO) drift, channel fading, and variable SNR.

1) *RadioML2016.10A Dataset:* The RadioML2016.10A dataset includes 220 000 signals with SNRs ranging from −20 to +18 dB and 11 types of modulations. Each signal sample is of size 2×128, including I and Q components.
2) *RadioML2018.01A Dataset:* The RadioML2018.01A dataset consists of 24 types of modulations (19 digital modulations and five analog modulations). The SNRs of radio signals range from −20 to +30 dB and the signal length is 1024.

We divide these two datasets into the training set, validation set, and test set with the ratio of 6:2:2 and then preprocess each signal sample into its power spectrum [41] and discrete wavelet transformation (DWT) [42]. The power spectrum is defined as

$$P(\omega) = \lim_{H \to \infty} \frac{|F_H(\omega)|^2}{2\pi H} \tag{8}$$

where $F_H(\omega)$ is the Fourier transform of signal $x(h)$ in the time range of $h \in [-H/2, H/2]$. DWT can be expressed as

$$\text{DWT}(p, q) = a_0^{-\frac{p}{2}} \int_{-\infty}^{+\infty} x(h)\psi(a_0^{-p}t - qb_0)dh \tag{9}$$

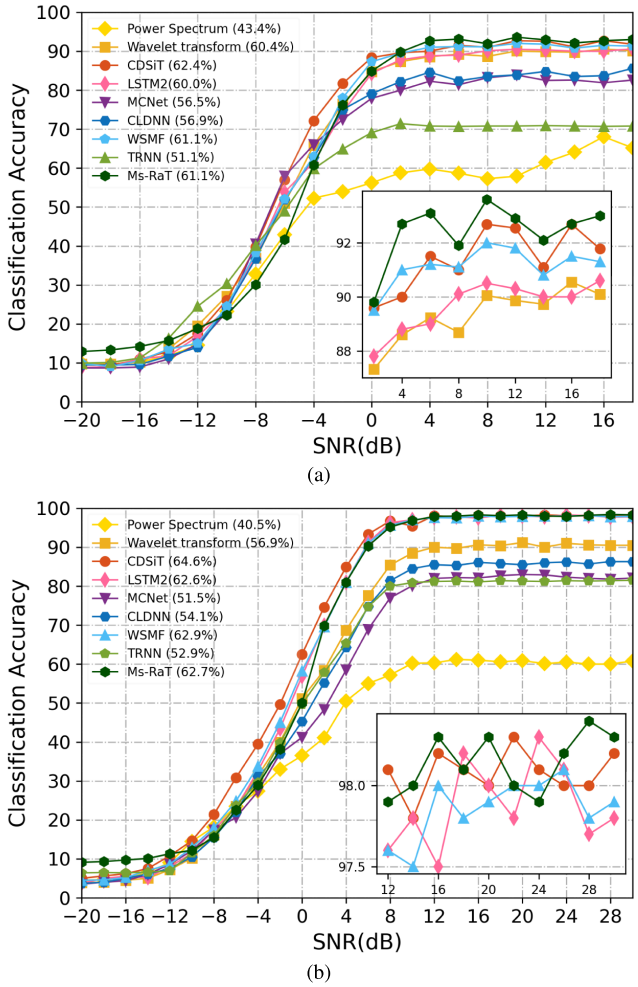

Fig. 2. Accuracy curves of different methods on (a) RadioML2016.10A dataset and (b) RadioML2018.01A dataset.

where $\psi(\cdot)$ is the wavelet basis function, $p, q = 0, \pm 1, \pm 2, \ldots$, and a_0, b_0, p, and q are the results of discretization of continuous wavelet transform scale parameter and translation parameter. In DWT, the signal is decomposed into two parts, the low-frequency component CA_k and the high-frequency component CD_k. In this brief, CA_k and CD_k are concatenated as inputs of the "wavelet transform" branch.

In the experiments, we use the logarithmic power spectrum and the first-level Harr transform coefficients as inputs of dual path for an efficient and simple realization. The experiment is repeated ten times under the same condition and the average results are taken for a comparison.

B. Performance of Cross-Domain SiT

In this section, we compare our proposed CDSiT with some state-of-the-art DNN-based AMC methods, including LSTM2 [20], MCNet [12], CLDNN [11], WSMF [31], TRNN [23], and Ms-RaT [32]. CDSiT and WSMF have dual inputs, while the others only have one. In addition, we compare our proposed CDSiT with SiT trained on wavelet transform data, SiT trained on power spectrum data trained on wavelet transform data, and SiT trained on power spectrum data, to investigate the performance of cross-domain learning. Fig. 2 shows that the overall accuracy (OA) of the RadioML2016.10A and RadioML2018.01A datasets by the nine methods is shown in Fig. 2(a) and (b), respectively. From the results, we can observe that CDSiT has competitive performance in terms of OA on the two

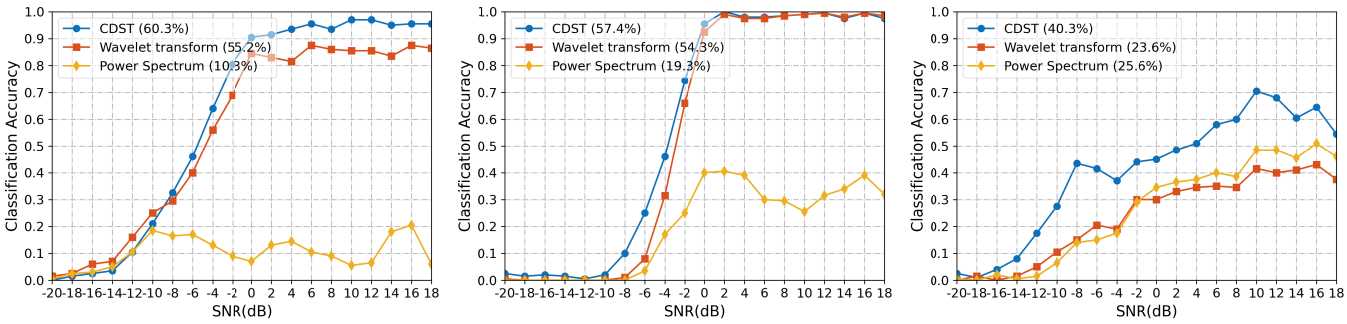

Fig. 3. Classification accuracy of the proposed model on RadioML2016.10A dataset with different SNRs. The input of CDSiT is wavelet transform coefficients and power spectrum of the signal. In the legend, "wavelet transform" is SiT trained on wavelet coefficients of the signal, and "power spectrum" is SiT trained on power spectrum of the signal. (a) QAM16. (b) QPSK. (c) WBFM.

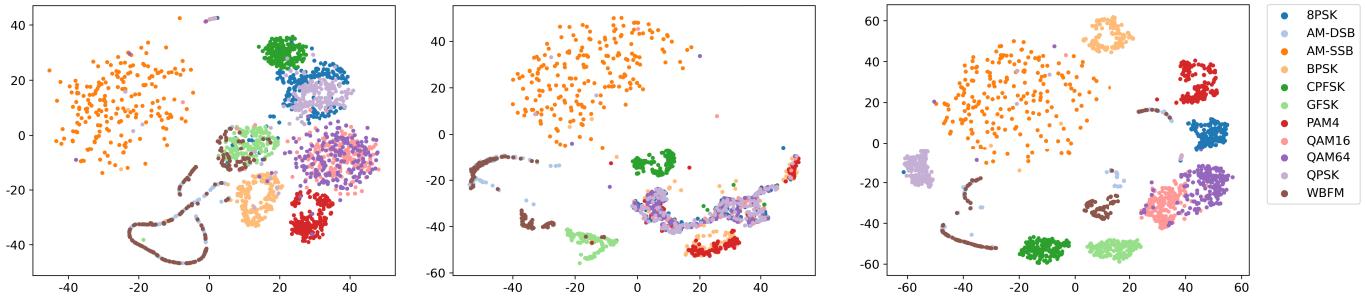

Fig. 4. Visualization of the input and output features of SFB on the RadioML2016.10A dataset. (a) Input of the SFB on wavelet transform branch. (b) Input of the SFB on power spectrum branch. (c) Fused output of the SFB.

TABLE I
COMPARISON BETWEEN CDSIT AND TRADITIONAL METHODS

Method	Input of model		Accuracy/%	Time/μs
	Power spectrum	Wavelet transform		
SVM	✓		12.0	2050.27
		✓	16.1	2645.86
KNN	✓		29.0	185.55
		✓	29.9	188.49
DTree	✓		27.0	0.73
		✓	18.2	0.73
XGBoost	✓		39.0	1.13
		✓	35.1	1.16
SiT	✓		43.4	4119.77
		✓	60.4	4121.84
CDSiT	✓	✓	62.4	4906.93

datasets. In Fig. 2(a), CDSiT has an improvement of 1.3%–7.1% over the other methods. In particular, CDSiT has a large increase of OA at SNRs from −8 to 0 dB, which validates its robustness and stability in complex environments. Meanwhile, in Fig. 2(b), CDSiT has an improvement of 1.9%–13.1% than the others. Similarly, compared with its counterparts, CDSiT has an obvious advantage from −10 to 8 dB. This demonstrates the superiority of CDSiT to other methods. Furthermore, we can also see that CDSiT outperforms SiTs trained solely on wavelet transform and power spectrum, which validates the cross-domain learning capability of CDSiT.

We then compare the performance of several traditional machine learning methods, including support vector machine (SVM) [6], K-nearest neighbors (KNNs) [7], decision trees (DTrees) [8], and XGBoost [9], with the deep learning methods SiT and CDSiT. Power spectrum and wavelet transform are separately or cascaded as inputs of the models. Table I shows the OA and consumed time of these methods. From the results, we can observe that compared with DNN-based methods, the traditional methods have faster inference for their shallow structure. For most of the traditional methods, using the power spectrum as input led to higher accuracy for traditional methods. However, deep learning methods are good at capturing complex nonlinear relationships and achieve significantly higher accuracy, up to twice as high as that of traditional methods. In terms of both performance and complexity, CDSiT shows superior practical applicability.

Table II compares the results of some combinations of multimodal description from different domains, such as I/Q components, constellation diagram, power spectrum, and wavelet transform. Experiments are first conducted with two-modal inputs. Because the constellation diagram is susceptible to noise, it can be observed that the combinations, including constellation diagram, present relatively poor performance compared to the other combinations (2.2%–7.7% lower than the best results). On the contrary, the results of the combinations, including wavelet transform, are significantly higher, especially for the combination of wavelet transform with power spectrum, which confirms the feasibility of our used combination in CDSiT. In addition, experiments with more signal descriptions as inputs are also conducted. Because the designed SFB can only fuse two modalities, the CDSiT can be modified to deal with

TABLE II
FUSION RESULT OF INPUT FROM EACH FIELD

Domain 1	Domain 2	Domain 3	Domain 4	Accuracy/%	Time/ms
I/Q	Constellation diagram	-	-	58.9	4.88
I/Q	Power spectrum	-	-	60.4	4.92
I/Q	Wavelet transform	-	-	60.1	4.90
Constellation diagram	Power spectrum	-	-	54.0	**4.87**
Constellation diagram	Wavelet transform	-	-	59.5	**4.87**
Power spectrum	Wavelet transform	-	-	**62.4**	4.91
I/Q	Constellation diagram	Power spectrum	-	56.5	10.87
I/Q	Constellation diagram	Wavelet transform	-	59.5	10.85
I/Q	Power spectrum	Wavelet transform	-	61.5	10.96
Constellation diagram	Power spectrum	Wavelet transform	-	61.3	10.85
I/Q	Constellation diagram	Power spectrum	Wavelet transform	60.9	91.82

TABLE III
EFFECT OF DWT RESULTS CA_k AND CD_k ON OA

	RadioML2016.10A	RadioML2018.01A
I/Q	60.4%	61.3%
$[CA_k, CD_k]$	**62.4**%	**64.6**%
CA_k	60.9%	62.7%
$R(CA_k)$	60.3%	61.8%
$R(CA_k, filter(CD_k)$	60.4%	62.1%

*$[a, b]$ donates concatenate a and b. $R(a)$ indicates that a will be reconfigured. And $filter(\cdot)$ is the smoothing filter

n-modality inputs. In this case, there are possible C_n^2 fused models, and the learned features by separate models can be averaged for AMC. In Table II, we show the OA and inference time for $n = 3$ and $n = 4$. However, due to the representation capabilities of different signal descriptions and increased model complexity, various high-dimensional inputs have different performance improvements. This suggests that it is necessary to select complementary multimodal descriptions for classification. Simultaneously, as more modalities are integrated, the inference time on a single sample also increases, along with a slowdown improvement of model performance.

Furthermore, because DWT can generate two sets of coefficients: CA_k and CD_k, various combinations of them will yield different results of CDSiT. Table III shows the results produced by different combinations of CA_k and CD_k, where $[a, b]$ indicates that a and b are concatenated to a feature whose length is the same as that of original signal, $R(a)$ is the mapping that reconstructs a back to the length of the original signal, and filter$(\cdot)$ is a smoothing filter that experimentally takes a Gaussian filter with mean 0 and variance 1. From the results, we can see that higher precision can be achieved in the case of concatenating CA_k and CD_k since it well preserves high- and low-frequency information of the signal.

We further conduct experiments to compare the results of two inputs with a single input for three categories in the RadioML2016.10A dataset. The OA curves of CDSiT and two single-domain models for QAM16, QPSK, and WBFM are plotted in Fig. 3(a)–(c), respectively. From the results, we can observe that CDSiT has an accuracy improvement from 4.5% to 49.4% over the single-domain methods for the classification of QAM16. Similar improvements can be observed on the QPSK and WBFM. In addition, from the results in Fig. 3(a) and (b), we can observe that the wavelet transform has a significant positive influence on CDSiT despite very low OA by power spectrum. From the results in Fig. 3(c), we can see that the power spectrum can better classify WBFM. This also validates the complementarity of power spectrum and wavelet transform.

In addition, Fig. 4 shows the visualization of the input and output features of SFB on the RadioML2016.10A dataset. T-SNE [43] is used to visualize the network features and to investigate how SFB changes the representation of inputs as they pass through the fusion layer. From the results in Fig. 4(a), we can observe that the wavelet transform branch can learn a good feature distribution for BPSK, AM-SSB, BPSK, CPFSK, PAM4, and QPSK, which suggests that SFB can learn compact intraclass and separate interclass features. In Fig. 4(b), most of the modulation classes have small interclass feature distances and also have a large intersection of feature distributions. This illustrates that the features in the power spectrum branches could not distinguish most of the modulation classes but are discriminative for AM-SSB, CPFSK, GFSK, and WBFM. Fig. 4(c) shows the output features of SFB, from which we can observe that SFB learns a good distribution of features for the modulation types and reduces intraclass distances while increasing interclass distances. In addition, by combining the advantages of the two paths, SFB improves the recognition of some hard classes, such as QAM16 and QAM64.

We also perform ablation experiments for SE, SFB, and TCN block on the RadioML 2016.10A dataset. Without these three modules, the model structure is the same as that of CDSiT, and the inputs of the dual path remain unchanged. The input embedding is the division of the input into multiple patches. In addition, the fusion part connects the two feature maps, and the classification head is a linear classifier. The remaining parameter settings are the same as those of the CDSiT model. The experimental results are shown in Table IV, from which we can see that the three modules are beneficial for modulation classification. When the SE module and the SBF block are used, the OA can be improved by 14.7% and 3.7%, respectively. In

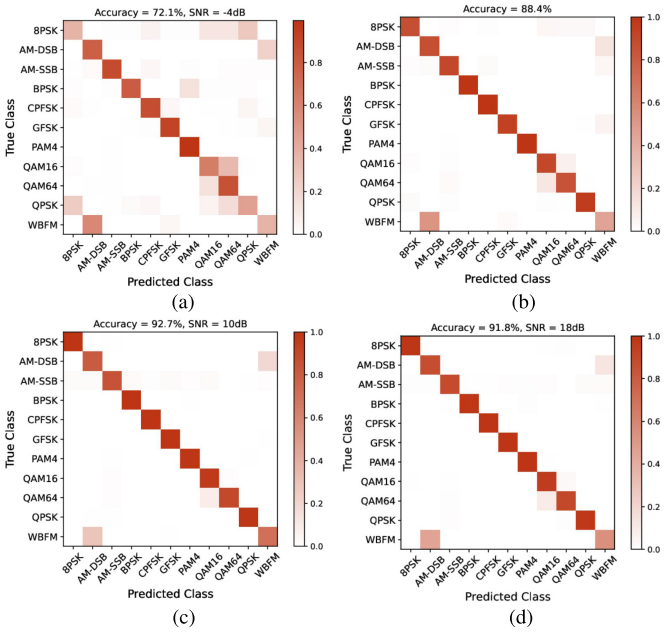

Fig. 5. Confusion matrix of the proposed CDSiT on RadioML2016.10A dataset. (a) SNR $= -4$ dB. (b) SNR $= 0$ dB. (c) SNR $= 10$ dB. (d) SNR $= 18$ dB.

TABLE IV
ABLATION EXPERIMENTS FOR SE, SFB BLOCK, AND TCN BLOCK

Signal Embedding	SFB block	TCN block	Accuracy/%
			42.9
✓			57.6
✓	✓		61.3
✓	✓	✓	62.4

particular, SE, which has induction bias and local spatial information, also plays an important role in model performance improvement.

Finally, we show the classification results of CDSiT with different SNRs. The confusion matrices of the classification results with SNRs -4, 0, $+10$, and $+18$ dB are shown in Fig. 5(a)–(c). From them, we can observe that all modulation types are well distinguishable, except that WBFM is easily confused with AM-DSB. Fig. 4(c) shows that WBFM and AM-DSB have very similar feature distributions, which will lead to misclassification. Because wavelet transform separates the high- and low-frequency components of the raw signal, CDSiT is robust to Gaussian noise. When it comes to identifying high-order modulation modes, CDSiT also delivers excellent results for efficient fusion of signal features across domains.

IV. CONCLUSION

In this brief, we propose a novel CDSiT model for AMC under the paradigm of multimodality learning. The designed SE, signal transform, SFB, and temporal convolution head are beneficial for modeling the latent correlation in multimodal descriptions, to learn more comprehensive and discriminative signal features for AMC. Compared with the available works, CDSiT yields competitive results in AMC by a tentative fusion of wavelet coefficients and power spectrum. Furthermore, CDSiT is scalable and can integrate information from multiple domains in a simple way. The complementarity of more signal descriptions will be investigated in future work.

REFERENCES

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

[2] F. Wang, O. A. Dobre, C. Chan, and J. Zhang, "Fold-based Kolmogorov–Smirnov modulation classifier," *IEEE Signal Process. Lett.*, vol. 23, no. 7, pp. 1003–1007, Jul. 2016.

[3] L. De Vito, S. Rapuano, and M. Villanacci, "Prototype of an automatic digital modulation classifier embedded in a real-time spectrum analyzer," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 10, pp. 2639–2651, Oct. 2010.

[4] M. Liu, N. Qu, J. Tang, Y. Chen, H. Song, and F. Gong, "Signal estimation in cognitive satellite networks for satellite-based industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2062–2071, Mar. 2021.

[5] H. Mustafa and M. Doroslovacki, "Digital modulation recognition using support vector machine classifier," in *Proc. 38th Asilomar Conf. Signals, Syst. Comput.*, 2004, pp. 2238–2242.

[6] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. their Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

[7] J. Laaksonen and E. Oja, "Classification with learning K-nearest neighbors," in *Proc. Int. Conf. Neural Netw.*, 1996, pp. 1480–1483.

[8] Y.-Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, pp. 130–135, Apr. 2015.

[9] S. Parui, A. K. Roshan Bajiya, D. Samanta, and N. Chakravorty, "Emotion recognition from EEG signal using XGBoost algorithm," in *Proc. IEEE 16th India Council Int. Conf. (INDICON)*, Dec. 2019, pp. 1–4.

[10] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 168–179, Feb. 2018.

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

[12] T. Huynh-The, C.-H. Hua, Q.-V. Pham, and D.-S. Kim, "MCNet: An efficient CNN architecture for robust automatic modulation classification," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 811–815, Apr. 2020.

[13] S. Peng et al., "Modulation classification based on signal constellation diagrams and deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 718–727, Mar. 2019.

[14] M. Liu, G. Liao, N. Zhao, H. Song, and F. Gong, "Data-driven deep learning for signal classification in industrial cognitive radio networks," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3412–3421, May 2021.

[15] Y. Mao, M.-L. Zhu, T. Sun, Y.-Y. Dong, and C.-X. Dong, "Automatic modulation classification based on SNR estimation via two-stage convolutional neural networks," in *Proc. 6th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2021, pp. 294–298.

[16] Y. Mao, Y. Y. Dong, T. Sun, X. Rao, and C. X. Dong, "Attentive Siamese networks for automatic modulation classification based on multitiming constellation diagrams," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5988–6002, Dec. 2021.

[17] W. Xie, S. Hu, C. Yu, P. Zhu, X. Peng, and J. Ouyang, "Deep learning in digital modulation recognition using high order cumulants," *IEEE Access*, vol. 7, pp. 63760–63766, 2019.

[18] A. Fehske, J. Gaeddert, and J. H. Reed, "A new approach to signal classification using spectral correlation and neural networks," in *Proc. 1st IEEE Int. Symp. New Frontiers Dyn. Spectr. Access Netw.*, Nov. 2005, pp. 144–150.

[19] M. Liu, K. Yang, N. Zhao, Y. Chen, H. Song, and F. Gong, "Intelligent signal classification in industrial distributed wireless sensor networks based industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 4946–4956, Jul. 2021.

[20] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Trans. Cognit. Commun. Netw.*, vol. 4, no. 3, pp. 433–445, Sep. 2018.

[21] S. Li, F. Li, S. Tang, and F. Luo, "Heart sounds classification based on feature fusion using lightweight neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[22] F. Liu, Z. Zhang, and R. Zhou, "Automatic modulation recognition based on CNN and GRU," *Tsinghua Sci. Technol.*, vol. 27, no. 2, pp. 422–431, Apr. 2022.

[23] J. Cai, F. Gan, X. Cao, and W. Liu, "Signal modulation classification based on the transformer network," *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 3, pp. 1348–1357, Sep. 2022.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[25] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–11.

[26] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[27] S. Alfasly, C. K. Chui, Q. Jiang, J. Lu, and C. Xu, "An effective video transformer with synchronized spatiotemporal and spatial self-attention for action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, Jul. 2022.

[28] X. Zhao et al., "Fractional Fourier image transformer for multimodal remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, Jul. 2022.

[29] C. Li, Z. Zhang, X. Zhang, G. Huang, Y. Liu, and X. Chen, "EEG-based emotion recognition via transformer neural architecture search," *IEEE Trans. Ind. Informat.*, vol. 19, no. 4, pp. 6016–6025, Apr. 2023.

[30] A. Moradzadeh, H. Moayyed, B. Mohammadi-Ivatloo, G. B. Gharehpetian, and A. P. Aguiar, "Turn-to-turn short circuit fault localization in transformer winding via image processing and deep learning method," *IEEE Trans. Ind. Informat.*, vol. 18, no. 7, pp. 4417–4426, Jul. 2022.

[31] P. Qi, X. Zhou, S. Zheng, and Z. Li, "Automatic modulation classification based on deep residual networks with multimodal information," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 1, pp. 21–33, Mar. 2021.

[32] Q. Zheng, P. Zhao, H. Wang, A. Elhanashi, and S. Saponara, "Fine-grained modulation classification using multi-scale radio transformer with dual-channel representation," *IEEE Commun. Lett.*, vol. 26, no. 6, pp. 1298–1302, Jun. 2022.

[33] S. Sahay, E. Okur, S. H. Kumar, and L. Nachman, "Low rank fusion based transformers for multimodal sequences," in *Proc. 2nd Grand-Challenge Workshop Multimodal Lang.*, 2020, pp. 29–34.

[34] M. H. Shah and X. Dang, "Novel feature selection method using Bhattacharyya distance for neural networks based automatic modulation classification," *IEEE Signal Process. Lett.*, vol. 27, pp. 106–110, 2020.

[35] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 10944–10956.

[36] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv:2104.05704*.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[38] Y. Hu, Y. Chen, W. Yang, L. He, and H. Huang, "Hierarchic temporal convolutional network with cross-domain encoder for music source separation," *IEEE Signal Process. Lett.*, vol. 29, pp. 1517–1521, 2022.

[39] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.

[40] T. J. O'shea and N. West, "Radio machine learning dataset generation with GNU radio," in *Proc. GNU Radio Conf.*, Sep. 2016, vol. 1, no. 1.

[41] M. Kulin, T. Kazaz, I. Moerman, and E. De Poorter, "End-to-end learning from spectrum data: A deep learning approach for wireless signal identification in spectrum monitoring applications," *IEEE Access*, vol. 6, pp. 18484–18501, 2018.

[42] X.-Y. Wang and H. Zhao, "A novel synchronization invariant audio watermarking scheme based on DWT and DCT," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4835–4840, Dec. 2006.

[43] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 8, pp. 2579–2605, Aug. 2008.