Fine-grained Modulation Classification Using Multi-scale Radio Transformer with Dual-channel Representation

Qinghe Zheng, *Student Member, IEEE*, Penghui Zhao, Hongjun Wang, *Member, IEEE*, Abdussalam Elhanashi, *Student Member*, *IEEE*, and Sergio Saponara, *Senior Member, IEEE*

***Abstract*—Automatic modulation classification (AMC) plays a critical role in both civilian and military applications. In this letter, we propose a multi-scale radio transformer (Ms-RaT) with dual-channel representation for fine-grained modulation classification (FMC). In Ms-RaT, a dual-channel representation (DcR) of radio signals is designed to help the model learn discriminative features by converging multi-modality information, including frequency, amplitude, and phase. During the learning process, multi-scale analysis is introduced into the model to form the tighter decision boundary. Finally, extensive simulation results demonstrate that Ms-RaT can achieve superior modulation classification accuracy with the similar or lower computational complexity than existing state-of-the-art deep learning methods. Through ablation studies, we also validate the effectiveness of DcR and multi-scale analysis in Ms-RaT.**

***Index Terms*—5G communication, cognitive radio, fine-grained modulation classification (FMC), transformer, multi-scale analysis.**

I. Introduction

As the fundamental step of dynamic spectrum access (DSA) technique, automatic modulation classification (AMC) has become the cornerstone of fifth-generation (5G) and beyond 5G (B5G) wireless communications. The base station can select an appropriate modulation scheme according to the channel state so as not to exceed the given block error rate [1]. At present, to improve the signal transmission rate and spectrum efficiency, various higher-order modulation schemes have been applied to communication systems. It is necessary to develop fine-grained modulation classification (FMC) methods for distinguishing similar modulation schemes of different orders.

Compared with traditional AMC tasks, FMC aims to identify intra-class modulation schemes which are more challenging to characterize, such as multiple phase shift keying (MPSK) and multiple quadrature amplitude modulation (MQAM). Smaller inter-class and larger intra-class distances of these modulated signals are detrimental to the learning of the decision boundary of models. Although numerous studies have been developed for AMC, the FMC performance is still unsatisfactory and has not attracted enough attention, *e.g.* distinguishing between 16QAM and 64QAM [2,3]. The poor accuracy of high-order modulation classification failed to inspire follow-up research.

Commonly used AMC techniques are usually fall into two categories: likelihood based (LB) methods and feature based (FB) methods. By maximizing the likelihood probability under various assumptions such as Kolmogorov-Smirnov test (KST) [4] and average likelihood ratio test (ALRT) [5], LB methods determine the modulation scheme of received signals. Although LB methods guarantee optimal classification results in theory, they suffer from high computational complexity and require the prior knowledge of channel parameters, which is not applicable in practical applications. In FB methods, features like cumulant [6] are extracted and fed into classifiers to obtain classification results. However, the bias in the estimation of the fourth-order cumulant becomes more obvious as the order of the modulation scheme increases. Further, it is demanding to design suitable discriminative features for FMC. On the other hand, traditional classifiers (*e.g.* support vector machine) cannot cope with more compact feature distributions in FMC task.

Recently, deep learning has made remarkable achievements in pattern recognition and increasingly developed for AMC. Zeng *et al.* [7] proposed a convolutional neural network (CNN) based AMC framework driven by spectrogram. Chen *et al.* [8] designed a single-layer long short-term memory (LSTM) model based on the attention mechanism. Liu *et al.* [9] introduced an AMC method through exploiting graph convolutional network (GCN). However, most existing deep learning methods only utilized the monomodal information from a single description dimension, such as in-phase/quadrature (I/Q) temporal series or power spectrum in the frequency domain. Other modes of radio signals, such as instantaneous frequency and phase spectrum, have not been well explored.

In this letter, we propose an FMC method using multi-scale radio transformer with dual-channel representation. Extensive simulations covering 16 fine-grained modulation schemes are performed to validate the effectiveness of the proposed model. It works well for high-order modulation schemes even at lower SNRs. Main contributions are summarized as follows.

- The transformer network with better generalization ability is specifically developed for FMC.
- A dual-channel representation of radio signals is designed to help the model learn discriminative features.
- Multi-scale analysis is introduced into the model to help form a tighter decision boundary.

The rest of the letter is organized as follows. Section II states the signal model and FMC problem. Section III introduces the proposed multi-scale radio transformer. The simulation results are discussed in Section IV. Finally, we conclude our work and

This work was supported by National Key R&D Program of China under Grant 2018YFF01014304 and 2012YQ20022407, and Shandong Provincial Natural Science Foundation under Grant ZR2019ZD01, and. (*Corresponding author: Hongjun Wang.*)

Qinghe Zheng, Penghui Zhao, and Hongjun Wang are with the School of Information Science and Engineering, Shandong University, Qingdao 266237, China (e-mail: 15005414319@163.com; 201812603@mail.sdu.edu; hjw@sdu.edu.cn).

Abdussalam Elhanashi and Sergio Saponara are with the Department of Information Engineering, University of Pisa, Pisa 56122, Italy (email: a.elhanashi@studenti.unipi.it; sergio.saponara@unipi.it)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2022.3145647, IEEE Communications Letters

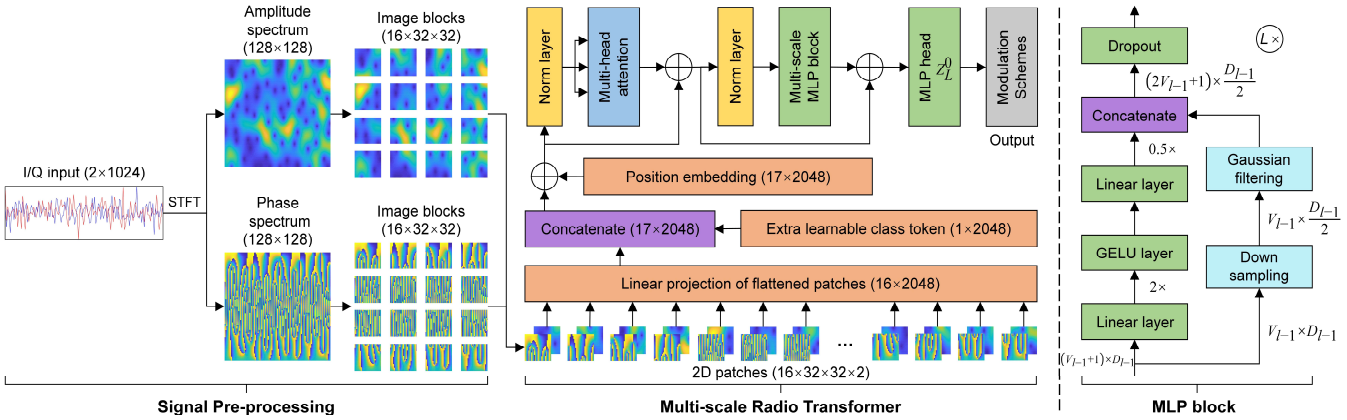

Fig. 1. Specific structure of multi-scale radio transformer with dual-channel representation.

future directions in Section V.

II. Problem Statement

Assume that the received baseband signal is sampled at the symbol rate from pulses satisfying the Nyquist limit, it can be expressed as

$$r_i(n) = \alpha_n e^{j(\Delta\varphi + \pi\Delta f n/N)} s_{c_i}(n) + w(n),\ n = 0,\ 1,\ \ldots,\ N-1 \quad (1)$$

where α_n is the time-varying channel amplitude gain following the Rayleigh distribution at the range of (0, 1]. $\Delta\varphi$ represents the carrier phase offset caused by propagation delay and initial carrier phase, and it obeys the Union distribution, *i.e.*, $\Delta\varphi \sim U(0, \pi/16]$. $\Delta f = 0.04$ denotes the normalized carrier frequency offset. $s_c(n)$ stands for the n-th transmitted symbol drawn from constellations of c-th modulation scheme, in which $c_i = 1, 2, \ldots, C$ and C is the total number of candidate modulation schemes. w is the additive white Gaussian noise (AWGN) with mean 0 and variance $2\sigma_w^2$. If constellations have unit power, the received signal-to-noise ratio (SNR) can be defined as $\gamma = \frac{1}{N}\sum_{n=0}^{N-1}\alpha_n^2 \Big/ 2\sigma_w^2$.

For the baseband radio signal r_i, our goal is to maximize the probability of correctly identifying its modulation scheme by adjusting model parameters, as given by

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathrm{P}(\mathbf{F}_{\boldsymbol{\theta}}(r_i) = [y_1, \ldots y_{c_i}, \ldots y_C] \,|\, \mathbf{y}_i = [y_1, \ldots y_{c_i}, \ldots y_C]) \quad (2)$$

where $\mathbf{F}_{\boldsymbol{\theta}}(\cdot)$ represents the machine learning classifier and $\boldsymbol{\theta}$ is learnable parameters. $\mathbf{y}_i$ denotes the ground-truth label of i-th received signal, in which y_{c_i} is 1 and the others are 0.

III. Multi-Scale Radio Transformer

To facilitate learning discriminative and rich representations from signals, we propose a multi-scale radio transformer (Ms-RaT) with dual-channel representation (DcR) for efficient FMC. The specific structure of Ms-RaT is shown in Fig. 1, and details are introduced as follows.

A. Dual-channel Representation

Since it is challenging for deep learning models to learn strict

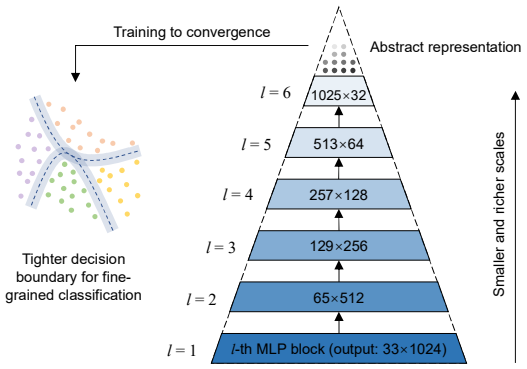

Fig. 2. Approximate multi-scale pyramid in Ms-RaT.

Fourier transform (FT), the spectrum analysis of signals is hard [16]. Therefore, the signal r comprised of I component r^I and Q component r^Q is first transformed into the frequency domain to obtain its amplitude spectrum (AS) κ and phase spectrum (PS) ψ, as given by

$$\begin{cases} \kappa = \sqrt{\Re^2(R) + \aleph^2(R)} \\ \psi = \arctan(\aleph(R)/\Re(R)) \end{cases} \quad (3)$$

where $\Re$ and $\aleph$ represent real and imaginary parts, respectively. R is the short-term Fourier transform (STFT) result and can be calculated by

$$R(p,q) = \sum_{n=0}^{N-1} r(n)\tau(n - pH)e^{\frac{-j2\pi qn}{W}} \quad (4)$$

where p and q represent discrete varying time and frequency, respectively. The sampling point is the same as window size W. $\tau(\cdot)$ denotes the window function like Hamming whose window size W and hop size H are set to 128 and 7 respectively, and can be defined as

$$\tau(n) = \begin{cases} 0.53836 - 0.46164\cos\left(\dfrac{2\pi n}{W-1}\right), & 0 \le n \le W-1 \\ 0, & else \end{cases} \quad (5)$$

It is necessary to pay attention to the tradeoff between the time

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2022.3145647, IEEE Communications Letters

and frequency resolution.

Then amplitude and phase spectra are connected in parallel as DcR, *i.e.*, $x = \{\kappa; \psi\} \in \mathbb{R}^{\frac{N-W}{H} \times W \times 2}$. As a preprocessed input, DcR helps to establish the robust mapping ($\zeta: r \rightarrow s$) from radio signals to their modulation schemes, considering all modulation information is reflected in amplitude and phase.

B. Radio Transformer

In this part, the radio transformer $\mathbf{F}_\theta(\cdot)$ is proposed to learn features from DcR. Firstly, the input is reshaped into a sequence $\hat{x} = [\hat{x}^1; \hat{x}^2; ...; \hat{x}^{V_0}] \in \mathbb{R}^{V_0 \times (2K^2)}$ of flattened 2D spectrum patches where $V_0 = \frac{(N-W)W}{HK^2} = 16$ represents the number of patches and $K \times K \times 2$ (32×32×2) is the resolution of each patch. Next, a latent vector $\eta_0 \in \mathbb{R}^{2K^2 \times D_0}$ (D_0 = 2048) with trainable linear projection is used for patch embedding as

$$z_0 = [\eta_0^{class}; \hat{x}^1\eta_0; \hat{x}^2\eta_0; ...; \hat{x}^{V_0}\eta_0] + \eta_0^{pos} \in \mathbb{R}^{(V_0+1) \times D_0} \quad (6)$$

where η^{class} is an additional learnable class token embedded in the patch sequence. η^{pos} denotes the position embedding for preserving positional information and can be manually encoded as

$$\eta_0^{pos}(i, j) = \begin{cases} \sin(pos_i / 10000^{j/D_0}), & j \in [2, 4, ..., D_0] \\ \cos(pos_i / 10000^{j/D_0}), & j \in [1, 3, ..., D_0 - 1] \end{cases} \quad (7)$$

where $pos_i = i \in [0, 1, \ldots, V_0]$ is the position of i-th 2D spectrum patch in the sequence.

Then embedded patches are fed into the transformer encoder consisted of alternating layers of multi-headed self-attention (MSA) [10] and multi-layer perceptron (MLP) blocks (L ×). MSA is an extension of SA where $k = 8$ self-attention operations (or heads) are run in parallel, and their concatenated outputs are projected according to

$$\mathrm{MSA}(z_0) = [\mathrm{SA}_1(z_0); \mathrm{SA}_2(z_0); ...; \mathrm{SA}_k(z_0)]\boldsymbol{\theta}_{msa} \quad (8)$$

$$\mathrm{SA}(z_0) = \mathrm{softmax}(\mathbf{q}\mathbf{k}^{\mathrm{T}} / \sqrt{D_0/k})\mathbf{v} \quad (9)$$

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = z_0\boldsymbol{\theta}_{sa} \quad (10)$$

where projections $\boldsymbol{\theta}_{msa} \in \mathbb{R}^{D_0 \times D_0}$ and $\boldsymbol{\theta}_{sa} \in \mathbb{R}^{D_0 \times \frac{3D_0}{k}}$ are parameter matrices. The norm layer (NL) [11] and residual connection are applied before and after each block respectively, as given by

$$z_l = \mathrm{MLP}_l(\mathrm{NL}(\hat{z}_l)) + \hat{z}_l, \; l \in [1, 2, ..., L] \quad (11)$$

$$\hat{z}_l = \mathrm{MSA}(\mathrm{NL}(z_{l-1})) + z_{l-1} \quad (12)$$

$$\mathrm{NL}(z_{l-1}) = \gamma \frac{z_{l-1} - \mu_z}{\sigma_z} + \beta \quad (13)$$

where scale γ and bias vector β are both learnable parameters. μ_z and σ_z represent mean and standard deviation of elements in z, respectively.

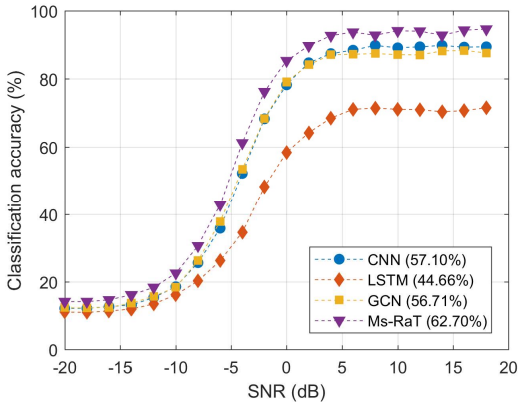

Fig. 3. Comparison of classification accuracy of various deep learning models.

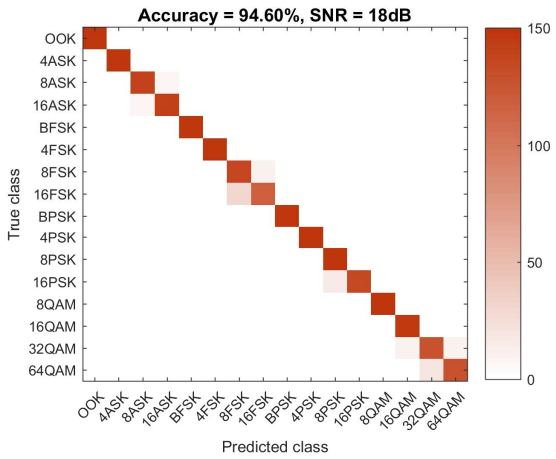

Fig. 4. The confusion matrix of Ms-RaT at +18 dB.

Finally, the output $z_L^0 = \eta_L^{class}$ of transformer encoder serves as the final representation $\mathrm{NL}(z_L^0)$ attached to a classification head implemented by an MLP with one hidden layer containing 256 neurons. All the model parameters can be updated through optimizing the cross-entropy loss using gradient descent.

C. Multi-scale Pyramid (MsP)

During the stacking of L = 6 MLP blocks, we construct the MsP to help extract more discriminative features and form the tighter decision boundary, as shown in Fig. 2. In MLP blocks, the down-sampling layer and Gaussian filtering are introduced to rescale feature patches (except for the class token) from the last block, as given by

$$\mathrm{MLP}_l(input) = \mathrm{Drop}[\mathrm{LL}_2(\mathrm{GELU}(\mathrm{LL}_1(input))); \; input(2:2:\mathrm{end}) * \delta_l] \quad (14)$$

where $\mathrm{LL}_1(\cdot)$ and $\mathrm{LL}_2(\cdot)$ are two linear layers using point-wise convolution, and Drop[·] represents the Dropout function [12] avoiding over-fitting problem. δ_l is Gaussian filters with size of 1 × 5 and variance of $\varepsilon_l = 0.8^l$. GELU(·) denotes the Gaussian error linear units [13] that perform nonlinear transformation, as calculated by

$$\mathrm{GELU}(input) = input \times \Phi(input) \quad (15)$$

where Φ represents the probability function of standard normal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2022.3145647, IEEE Communications Letters

TABLE I
INFERENCE SPEED COMPARISON OF A SERIES OF DEEP LEARNING MODELS

Models	Parameters (k)	test time (ms)	Accuracy (%)
CNN [7]	199	13.20	57.10
LSTM [8]	200	19.84	44.66
GCN [9]	437	9.80	56.71
Ms-RaT ($L = 4$)	440	12.45	61.39
Ms-RaT ($L = 5$)	524	14.14	62.44
Ms-RaT ($L = 6$)	610	17.30	62.70

distribution.

After each MLP block, the number of output feature patches is doubled and the length is half of the original, *i.e.*, $V_l = 2V_{l-1}$ and $D_l = 0.5D_{l-1}$. In other words, the scales of feature patches become richer as the number of learnable parameters increases. It is worth noting that two learnable vectors $\eta_{\mathrm{LL1}\_l} \in \mathbb{R}^{(V_{l-1}+1)\times 2D_{l-1}}$ and $\eta_{\mathrm{LL2}\_l} \in \mathbb{R}^{(V_{l-1}+1)\times 0.5D_{l-1}}$ have also become deformable due to the scale transformation in MsP. Since the dimensions of two learnable vectors η_{LL1_l} and η_{LL2_l} are reduced to half through an MSA layer, the total computational costs are similar to that of single-head self-attention with full dimension.

Compared with the original transformer [17], only MsP is introduced into the model structure, and its complexity as the convolution form is much lower than that of the peer-to-peer computation in linear layers. According to the analysis in [10], the SA layer connects all positions through a constant number $O(1)$ of sequential operations whereas a recurrent layer requires $O(V)$ sequential operations. The computational complexity of SA, convolutional, and recurrent layers are $O(V^2 \cdot D)$, $O(V \cdot D^2)$, and $O(S \cdot V \cdot D^2)$ respectively, in which S represents the kernel size. It can be deduced that the SA layer is faster than both convolutional and recurrent layers when the number of patches V is less than the representation dimension D.

IV. SIMULATION AND ANALYSIS

In this section, extensive simulations and comparisons are performed to validate FMC performance of Ms-RaT.

A. Simulation Setup

There are 16 fine-grained modulation schemes (*i.e.*, $C = 16$) are investigated in simulations, including {*OOK*, *4ASK*, *8ASK*, *16ASK*, *BFSK*, *4FSK*, *8FSK*, *16FSK*, *BPSK*, *4PSK*, *8PSK*, *16PSK*, *8QAM*, *16QAM*, *32QAM*, *64QAM*} with SNRs varying from −20 dB to +18 dB. A total of 320,000 radio signals are generated by the public data generator RadioML [14] with pre-defined symbol length $N = 1024$, of which 70%, 15%, and 15% are divided for training, validation, and test, respectively.

The model is first pre-trained on RadioML 2018.01a [14] and then fine-tuned on simulation data using Adam optimizer [15]. Compared with training from scratch, pre-training on a large-scale dataset can provide a better initialization position. Hyper-parameters including learning rate, weight decay, exponential decay, batch size, and Dropout are initialized to 0.01, 0.0005, 0.9, 64, and 0.5, respectively. Both training and testing of Ms-RaT are carried out on the workstation consisted of Intel Core i9-7900k CPU, NVIDIA TITAN Xp GPU, 32 GB memory, and 1 TB storage.

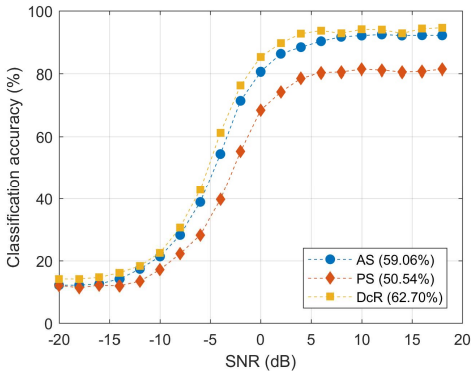

Fig. 5. The influence of DcR on FMC performance.

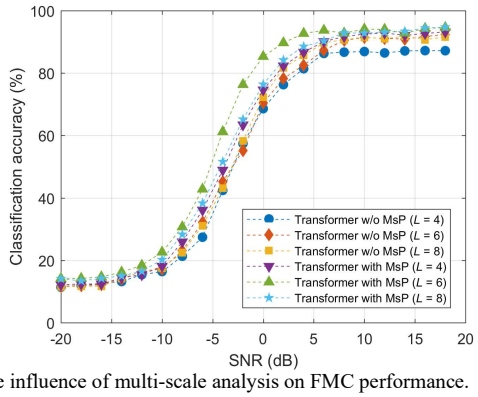

Fig. 6. The influence of multi-scale analysis on FMC performance.

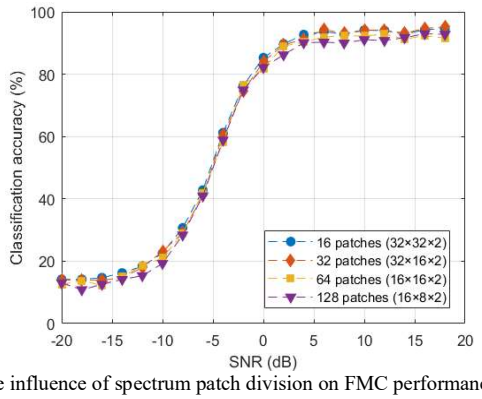

Fig. 7. The influence of spectrum patch division on FMC performance.

B. Classification Performance

In Fig. 3, we compare the classification accuracy of typical deep learning models to demonstrate the superiority of Ms-RaT, including CNN [7], LSTM [8], and GCN [9]. CNN is able to associate the context information in the spectrum image. LSTM is good at learning long-term dependencies. GCN can adapt to complex structure priors. It can be seen that Ms-RaT achieves the average classification accuracy of 5.60% to 18.04% higher than other models. In addition, the accuracy is improved by 5%-22% when SNR is between 0 dB and +18 dB. Ms-RaT shows significant advantages even at low SNRs. It should be noted that LSTM performs worst among these models and achieves the lowest average accuracy of 44.66%. Although the signal is sampled as time series, it is time independent and thus LSTM is not suitable for FMC.

To observe specific classification results of Ms-RaT, we give the confusion matrix at +18 dB SNR in Fig. 4. All modulation schemes can be well distinguished, except for a few of highest

order modulations. Ms-RaT exhibits the superior capability in identifying high-order modulation schemes due to its fusion of multimodal information and the usage of feature embedding. These properties enable transformer not only to be trained like conventional deep learning models fed with images, but also to more accurately establish the mapping between samples and categories with higher probability.

In Table I, we report the parameters and inference speeds of CNN [7], LSTM [8], GCN [9], and the proposed Ms-RaT with different number (L= 4, 5, 6) of MLP blocks. It can be seen that Ms-RaT performs better with similar or smaller test time due to the limited size of the DcR input. Although CNN has the least learnable parameters, its reasoning speed is slower than both LSTM and Ms-RaT (L = 4) due to the larger input spectrum size (200×200×3). In addition, the inference speed of Ms-RaT can be increased substantially with only small accuracy loss (<1.5%) by reducing the stacking of MLP blocks. This means that the model size can be adaptively determined according to the needs of practical applications.

C. Ablation Study

In this part, ablation study is performed to observe the effect of DcR as well as multi-scale analysis on FMC performance. In Fig. 5, we present the influence of DcR on FMC performance. Different types of inputs, including AS and PS, are compared to illustrate the effectiveness of DcR. According to the results, DcR achieves best accuracy at almost all SNRs. AS performs better than PS since more modulation information is reflected in the amplitude. Although there are only finite discrete values to characterize phase changes in digital modulations, PS can help overcome the limitations of AS.

In Fig. 6, we display the influence of multi-scale analysis on FMC performance. Different numbers of multi-scale pyramid modules are applied or removed in all MLP blocks to observe changes in classification accuracy. Compared to models with MsP removed, the transformers with multi-scale analysis earn better classification accuracy. As L increases, the multi-scale information of signals becomes richer and thus the accuracy can be improved. However, a larger L introduces more learnable parameters, which is more likely to cause over-fitting problems and affect the model robustness to varying SNRs.

Then we consider the impact of spectrum patch division on FMC performance. The classification accuracies driven by four patch division cases (16×32×32×2, 32×32×16×2, 64×16×16×2, and 128×16×8×2) are reported in Fig. 7. According to results, the size and number of patches have a negligible impact on the classification accuracy due to the patch embedding operation in Eq. (6). It is just the number of learnable parameters increases as the division scale becomes more refined, which affects the model optimization rather than generalization. In the practical applications, the local optimization efficiency of the model on the workstation is not critical, while the inference performance determines its availability.

V. Conclusion

In this letter, we propose a novel multi-scale radio transformer named Ms-RaT for FMC. Through simulation experiments and ablation studies, we illustrate the effectiveness and superiority of the model. It performs better than a series of typical deep learning models in the classification accuracy with the similar or faster inference efficiency, including CNN, LSTM, and GCN. To the best of our knowledge, this is the first attempt that DcR-driven radio transformer is developed for FMC. Furthermore, the multi-scale analysis has been proved to help improve FMC performance. In the future, the development and utilization of lightweight transformer is an important research direction for people to apply FMC techniques to 6G wireless communication networks. More complex noise and channel conditions will be also explored.

References

[1] F. Blanquez-Casado, M. D. C. A. Torres, and G. Gomez, "Link adaptation mechanisms based on logistic regression modeling," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 942–945, May 2019.

[2] Q. Zheng, P. Zhao, Y. Li, H. Wang, and Y. Yang, "Spectrum interference based two level data augmentation method in deep learning for automatic modulation classification," *Neural Comput. Appl.*, vol. 33, no. 13, 7723–7745, Nov. 2020.

[3] P. H. Qi, X. Y. Zhou, S. L. Zheng, and Z. Li, "Automatic modulation classification based on deep residual networks with multimodal information," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 21–33, Mar. 2020.

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

[5] W. Chung, "Sequential likelihood ratio test under incomplete signal model for spectrum sensing," *IEEE Trans. Wirel. Commun.*, vol. 12, no. 2, pp. 494–503, Feb. 2013.

[6] M. Abdelbar, W. H. Tranter, and T. Bose, "Cooperative cumulants-based modulation classification in distributed networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 3, pp. 446–461, Sept. 2018.

[7] Y. Zeng, M. Zhang, F. Han, Y. Gong, and J. Zhang, "Spectrum analysis and convolutional neural network for automatic modulation recognition," *IEEE Wirel. Commun. Lett.*, vol. 8, no. 3, pp. 929–932, Jun. 2019.

[8] Y. Chen, W. Shao, J. Liu, L. Yu, and Z. Qian, "Automatic modulation classification scheme based on LSTM with random erasing and attention mechanism," *IEEE Access*, vol. 8, pp. 154290–154300, Aug. 2020.

[9] Y. B. Liu, Y. Liu, and C. Yang, "Modulation recognition with graph convolutional network," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 5, pp. 624–627, May 2020.

[10] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, USA, Dec. 2017, pp. 5998–6008.

[11] R. Xiong *et al.*, "On layer normalization in the transformer architecture," in *Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, Jul. 2020, pp. 10524–10533.

[12] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[13] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–10.

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

[15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, California, US, May 2015, pp. 1–15.

[16] X. Wu, R. Tao, D. Hong, and Y. Wang, "The FrFT convolutional face: toward robust face recognition using the fractional Fourier transform and convolutional neural networks," *Sci. China-Inf. Sci.*, vol. 63, no.1, pp. 1–3, Oct. 2019.

[17] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, Mar. 2021, pp. 1–21.