Robust and Efficient Modulation Recognition with Pyramid Signal Transformer

He Su, Xinyi Fan
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, China
{suhe, xyxy}@njust.edu.cn

Huajun Liu*
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, China
liuhj@njust.edu.cn

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10001593

***Abstract*—A robust and efficient pyramid signal Transformer model, called SigFormer for automatic modulation recognition was proposed in this paper. In SigFormer, a pyramid Transformer architecture is introduced to encode the relationship between the internal features of modulated signals. Specifically, a dual-attention block composed of self-attention layer and scaling-attention layer is proposed for simultaneous global feature representation and noise resistance learning for modulated signals, and small-kernel convolution layers embedded to dual-attention block and feed-forward block is proposed for fine-grained modulation recognition as well. Experiments on RML2018.01a, RML2016.10a and RML2016.10b show that the SigFormer outperformed most other deep learning models on recognition accuracy, and it is more parameter-efficient than most other models and more robust on low signal-to-noise ratio (SNR) signals.**

***Index Terms*—Modulation recognition, Noise resistance learning, Dual-attention, Transformer**

I. Introduction

Rapidly recognition and understanding of the radio spectrum in an autonomous way is a key capability for spectrum interference monitoring, radio fault detection, dynamic spectrum access, opportunistic mesh networking, non-cooperative communication for 5G and beyond [1]. And identifying signal modulation mode under low signal-to-noise ratio (SNR) is a challenging problem in wireless communication.

Traditional automatic modulation recognition (AMR) methods, for instance likelihood-based methods [2] and feature-based methods [3, 4] usually depend on redundant prior knowledge and manually tuned parameters. Deep learning (DL) as a powerful tool has achieved great progress in the field of AMR in recent years. For instance, convolutional neural networks (CNNs) [5], recurrent neural networks (RNNs) [6], and long short-term memory (LSTM) [7, 8] have achieved better performance on this topic. However, classical DL methods usually fell into the local feature representation and lacked the global feature modeling capability for modulated signals, which would limit the recognition performance on most modulation recognition datasets.

Self-attention [9] and Transformer [10–12] as novel inventions in natural language processing (NLP) and computer vision (CV) communities have shown great potential to classify the radio signals [13]. However, original Transformers paid more attention on the self-similarity analysis from the view of long-range interaction, but ignored the local contextual features modeling which would misclassify the similar modulation types on large-scale datasets, and it is sensitive to noise and will degrade rapidly on low SNR signals because the original self-attention is usually attributed to the tensor matrix projection mechanism on the entire input space where noises are incorporated equally.

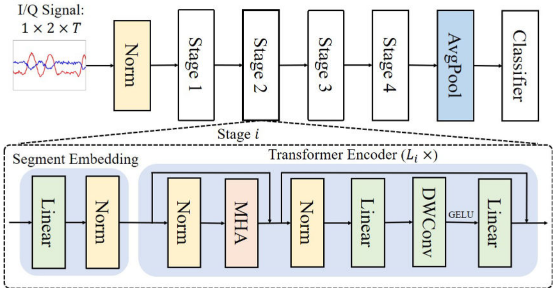


Fig. 1: The architecture of SigFormer.

In this paper, a robust signal Transformer model, called SigFormer, is built to identify the modulated signal modulations by simultaneous global-, local-feature representation and noise resistance learning in a pyramid Transformer architecture. Specifically, the main contribution of this work can be summarized as follows:

- Proposing an efficient and robust pyramid Transformer classifier to identify the signal modulation type;
- Proposing a dual-attention block with parallel self-attention and scaling-attention layers for simultaneous global feature extraction and noise resistance learning;
- Proposing efficient small-kernel convolutional layers embedding to enhance detailed local features representation for fine-grained modulation recognition.

II. Related Work

Since deep learning approaches have been introduced to modulation recognition tasks, the recognition accuracy has been improved greatly. For instance, O'Shea et al. [14] summarized the emerging applications of deep learning on radio signal processing and how to use GNU radio to generate

*Corresponding author

an open dataset of modulated signals with in-phase and quadrature (I/Q) information for modulation recognition. At the same time, Convolutional neural networks are applied for modulation recognition. Among them, the 1D-CNN [1] and 2D-CNN [5] are proposed for small-scale radio signal classification tasks. Benefiting from the translation invariance of convolutional layers and advanced CNN architectures, several CNN variants, such as SigNet [15], SCGNet [16] are applied to modulation recognition and achieved better recognition performance.

Meanwhile, the signal modulation recognition task is modeled as a time-series classification problem from the temporal feature representation as well. And several temporal representation models, such as RNN and LSTM [6, 8] etc. and spatio-temporal hybrid structures, such as CNN-LSTM [17], DCN-BiLSTM [7] and MCLDNN [18] etc., are comprehensively investigated for modulation recognition.

Above spatial, temporal or spatio-temporal hybrid modeling methods on modulated signals attend to extract signal's high dimensional discriminative features from local representation layers, i.e. convolutions in CNN, logic gates in LSTM, which is sensitive to signal noise. They lack the long-range feature modeling capability from the global view of modulated signal, which is significantly helpful for AMR on large-scale dataset and fine-grained signal identification scenario.

The self-attention dominated by tensor dot-product operations provides a novel mechanism to capture long-range interaction and analyze feature self-similarity from the global view, which has proven to be an efficient method for feature long-range interaction modeling. Self-attention augmented deep CNN [19] and LSTM [20] methods have been applied for signal modulation recognition. Very recent Transformer network, whose soul is the self-attention mechanism, has achieved amazing performance from NLP to CV [11] tasks, and it showed great potential to classify the radio signals [13]. But original Transformer models is also sensitive to noise especially on low SNR signals because it usually ignored the local neighbour smoothing idea and operations. In this work, we start with a pyramid Transformer model, and explore a robust signal modulation recognition method with simultaneous global feature representation and noise resistance learning mechanism.

III. Our Approach

A. Overall architecture

Our SigFormer model is built on dual-attention based Transformer architecture for signal modulation recognition, which adopts hierarchical structure to learn the internal representation of signal features with different SNR. The overall architecture of SigFormer is shown in Fig. 1, which consists of four stages feature encoding, and each stage includes a segment embedding layer and several Transformer encoder layers.

Firstly, the input modulated signal $\mathbf{X}_0 \in \mathbb{R}^{C_0 \times N_0 \times T_0}$ (C_0 is the feature dimension of each channel, N_0 means the number of channels, and T_0 is the length of signal in temporal domain) recorded by the I/Q sequence is pre-processed by an amplitude normalization on the sense of average power. Then the normalized signal is fed to a hierarchical Transformer model with four stages. In each stage, the input tensor $\mathbf{X}_0$ is fed to a segment embedding layer and several Transformer encoder layers, where the former divides the signal to several segments and encodes them to feature space, and simultaneously shrinks the signal length when necessary, and the latter extracts the internal discriminative features. Finally, a full connection layer maps the high-dimensional encoding features to a normalized confusion matrix of specific modulation classes for recognition tasks.

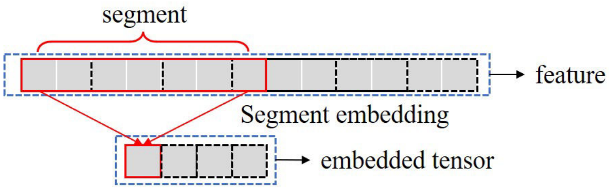


Fig. 2: Schematic example of segment embedding when $P_i = 7$ and $S_i = 2$.

B. Segment embedding

Segment embedding in signal Transformer aims to divide the signal to overlapping segments in temporal domain, and then linearly projection each segment to a higher-dimensional feature space, and when the step size between adjacent segments is greater than 1, it can play the role of time domain shrinkage, as shown in Fig. 2, which is similar with a patch embedding in vision Transformers [10, 11] who aims to divide the image into patches.

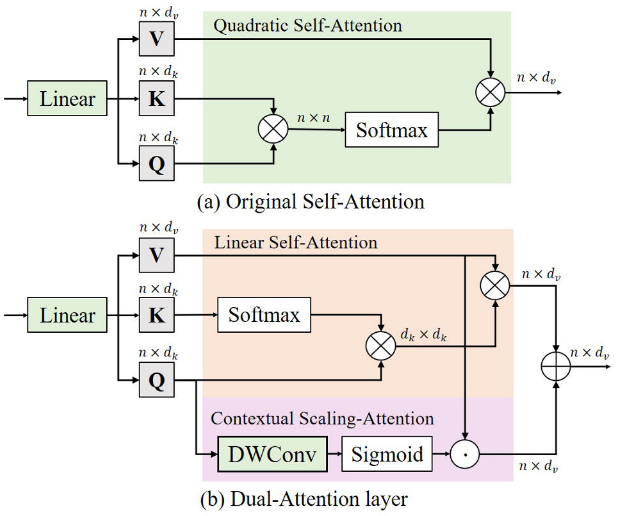
In the i-th stage, the segment length is denoted as P_i and the stride between segments is set as S_i. During segment embedding, the input feature $\mathbf{X}_{i-1} \in \mathbb{R}^{C_{i-1} \times N_{i-1} \times T_{i-1}}$ in hidden layers is firstly divided into $T_i = T_{i-1}/S_i$ overlapped segments, and then a learnable linear projection is multiplied with each segment and concatenated to obtain an embedded tensor $\mathbf{X}_i \in \mathbb{R}^{C_i \times N_i \times T_i}$. The overlapped segment embedding process can be formulated as:

$$\mathbf{X}_i = \mathrm{Concat}(\mathbf{x}_{i-1}^1\mathbf{E}, \mathbf{x}_{i-1}^2\mathbf{E}, \cdots, \mathbf{x}_{i-1}^{T_i}\mathbf{E}) \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{(N_{i-1}P_i \cdot C_{i-1}) \times C_i}$ is a learnable linear projection. When $S_i = 2$, the length of an input signal/feature will be reduced to a half, moreover N_i is set equal to the number of channels of the corresponding dataset samples in the first stage, so that $N_i = 1$ in the other stages.

C. Transformer encoder

There are L_i encoder blocks in the stage i of a Transformer encoder module, and each block consists of a multi-head attention (MHA) layer and a local feed-forward (LFF) layer. A batch normalization (BN) layer is inserted before each layer, and residual connection links before and after every layer.



(a) Original Self-Attention

(b) Dual-Attention layer

Fig. 3: The dual-attention block vs original self-attention.

Formally, the Transformer encoder module can be defined as:

$$\begin{aligned}\hat{\mathbf{X}}_i^{\ell-1} &= \mathrm{MHA}(\mathrm{BatchNorm}(\mathbf{X}_i^{\ell-1})) + \mathbf{X}_i^{\ell-1}\\ \mathbf{X}_i^{\ell} &= \mathrm{LFF}(\mathrm{BatchNorm}(\hat{\mathbf{X}}_i^{\ell-1})) + \hat{\mathbf{X}}_i^{\ell-1}\end{aligned} \quad (2)$$

where $\ell = 1, \cdots, L_i$, $\mathbf{X}_i^{\ell-1}$ and $\mathbf{X}_i^{\ell}$ denote the input and output feature of the ℓ-th encode block in the i-th stage, respectively. The dual-attention block, which is the workhorse of MHA, and the LFF block are defined in detail as follows.

- Dual-attention block. In the original self-attention [9, 13], the input tensor $\mathbf{X} \in \mathbb{R}^{n\times d}$ is linearly transformed into queries $\mathbf{Q} \in \mathbb{R}^{n\times d_k}$, keys $\mathbf{K} \in \mathbb{R}^{n\times d_k}$, and values $\mathbf{V} \in \mathbb{R}^{n\times d_v}$ respectively, and the original self-attention equation is defined as follows,

$$\mathrm{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (3)$$

where $\mathbf{Q}\mathbf{K}^T$ computes the $n \times n$ dimensional self-similarity between features on the global view with quadratic computation complexity and a Softmax function nonlinearizes them to highlight the content representation. However, the original version is sensitive to noise, the recognition accuracy will degrade quickly on lower SNR signals. And it lacks local contextual interaction modeling which is vital for fine-grained recognition on large-scale datasets.

To improve the recognition accuracy at low SNR, we propose a parallel contextual scaling-attention (CSA) layer, seen as Fig. 3(b), which uses the sigmoid nonlinearity function on local neighbouring information as a learnable smoother to resist the influence of signal noises. Meanwhile, a depth-wise convolution layer is introduced to harvest contextual information to provide local detailed feature representation. Our dual-attention block can be formulated as:

$$\mathrm{DA}(\cdot) = \mathrm{LSA}(\cdot) + \mathrm{CSA}(\cdot) \quad (4)$$

where $\mathrm{LSA}(\cdot)$ branch is a linear self-attention layer which works more efficiently based on channel-queried similarity analysis with computational complexity of $O(nd_k^2)$, which can be rewritten as

$$\mathrm{LSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \left(\frac{\mathbf{Q}^T\,\mathrm{Softmax}(\mathbf{K})}{\sqrt{n}}\right)\mathbf{V}^T \quad (5)$$

And the contextual scaling-attention branch $\mathrm{CSA}(\cdot)$ is a Sigmoid nonlinearization on a depth-wise convolutional features to smooth the local neighbouring features and reduce the noise of modulated signals, which is defined as

$$\mathrm{CSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Sigmoid}\left(\mathrm{DWConv}(\mathbf{Q})\right) \odot \mathbf{V} \quad (6)$$

where $\mathrm{DWConv}(\cdot)$ is a depth-wise convolution layer, $\mathrm{Sigmoid}(\cdot)$ is an active function and $\odot$ is the element-wise multiplication operation.

Finally, several single dual-attention "heads" are concatenated to form a multi-head attention (MHA) block, which can be defined as:

$$\mathrm{MHA}(\mathbf{X}) = \mathrm{Concat}\left(head_1, head_2, ..., head_h\right) \quad (7)$$

where $head_j \in \mathbb{R}^{n\times d_v}, j \in \{1, \cdots, h\}$ represents a single head tensor according to the (4).

- LFF block. The LFF block consisting of an expansion layer followed by a depth-wise convolution and a projection layer, which can be formulated as:

$$\mathrm{LFF}(\mathbf{X}_i^{\ell}) = \mathrm{Conv}(\mathrm{DWConv}(\mathrm{Conv}(\mathbf{X}_i^{\ell}))) \quad (8)$$

where a GELU activation layer closely follows the DWConv operator for nonlinearization. In the LFF, the first linear transformation layer expands the feature dimension by E_i times, and the third linear projection layer reduces the dimension by the same ratio.

D. Loss function

For training the model, a label smoothing cross-entropy [21] is defined as the loss function, which minimizes the expected value of the cross-entropy between the label smoothing ground truth q_k^{LS} and the model's inference results p_k as:

$$\mathcal{L} = -\sum_{k=1}^{K} q_k^{LS} log(p_k), \quad q_k^{LS} = q_k(1-\alpha) + \frac{\alpha}{K} \quad (9)$$

where q_k is "1" for the correct class and "0" for the rest, α is the label smoothing parameter and K is the total number of categories.

IV. Experiments

A. Datasets

Our SigFormer model is trained and evaluated on three signal modulation recognition data sets synthesized by GNU radio [29]: RML2016.10a, RML2016.10b [5, 14] and RML2018.01a [1]. On all datasets, the training set and the testing set are randomly divided by the ratio of 8:2.

- RML2016.10a contains a total of 220,000 signals for 11 public use modulation types collected in 2dB increments over a wide SNR range from -20dB to 18dB, each modulation contains 1,000 samples for each SNR. Each signal sample

TABLE I: Experiments on RML2016.10a, RML2016.10b and RML2018.01a

Dataset	Method	Avg. Accuracy↑	Highest Accuracy↑
RML2016.10a	GAF [22]	49.43%	79.30%
	MTF [22]	48.17%	76.60%
	Constellation [23]	44.92%	68.30%
	1D-ResNet [1]	57.74%	86.90%
	2D-CNN [5]	54.02%	80.50%
	LSTM [6]	57.20%	85.10%
	MCLDNN [18]	61.16%	92.70%
	DCN-BiLSTM [7]	62.10%	91.80%
	SigNet [15]	62.30%	91.30%
	Ours	**63.71%**	**93.60%**
RML2016.10b	1DCNN-PF [24]	57.79%	88.85%
	2D-CNN [5]	59.04%	85.51%
	CLDNN [25]	59.30%	85.53%
	CNN4 [26]	62.05%	90.25%
	GRU2 [27]	64.11%	93.55%
	LSTM [6]	64.16%	93.60%
	MCLDNN [18]	64.47%	94.02%
	MCformer [13]	65.04%	93.80%
	Ours	**65.77%**	**94.80%**
RML2018.01a	LSTM [6]	42.22%	68.24%
	ResNet [1]	49.39%	75.89%
	IRLNet [28]	60.07%	95.27%
	Ours	**63.96%**	**97.50%**

in both datasets has 128 complex floating-point I/Q sampling values.

- RML2016.10b contains 1,200,000 signals and is an extended version of RML2016.10a, which includes 10 types of modulated signals except AM-SSB.

- RML2018.01a contains 24 modulation types under 26 SNR (the step size is 2 dB in the range of - 20 dB to + 30 dB), each modulation type has 4,096 samples on each SNR, with a total of 2,555,904 signal samples. Each sample includes 1024 independent complex IQ samples.

B. Implementation details

For all experiments, AdamW [30] is adopted as the optimizer, and its momentum, weight decay are set to 0.9 and 1×10^{-4} respectively. The initial learning rate is set to 1×10^{-2} and decreases following the cosine schedule [31] with a minimum learning rate of 1×10^{-5}. On RML2016.10a, RML2016.10b and RML2018.01a, the batch size is set to 256, 1024, 512, and the training epochs are set to 30, 60, 60, respectively, where the first 10 epochs use the WarmUp [32] strategy. The hyper-parameter α for the label smoothing cross-entropy [21] loss function is set to 0.1. All the experiments are conducted on a NVIDIA TITAN Xp GPU with 12GB memory.

C. Comparison with other deep learning models

The overall performance of SigFormer can be concluded from the comparison experiments with the baseline models. The baseline models, such as GAF [22], MTF [22], 1D-ResNet [1], 2D-CNN [5], CLDNN [25], LSTM [6], MCLDNN [18], DCN-BiLSTM [7] and SigNet [15] are compared on the RML2016.10a, and baseline models, such as 1DCNN-PF [24], 2D-CNN [5], CNN4 [26], GRU2 [27], LSTM [6], MCLDNN [18] and MCformer [13] are compared on the RML2016.10b. And on RML2018.01a, our model are compared with ResNet [1], LSTM [6], IRLNet [28] etc. baseline models.

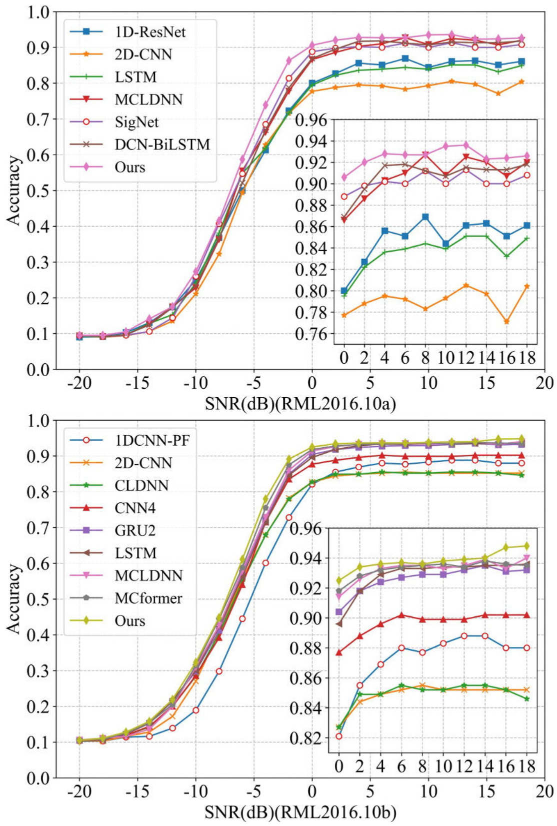


Fig. 4: Recognition accuracy on different SNR of SigFormer and other models on RML2016.10a and RML2016.10b.

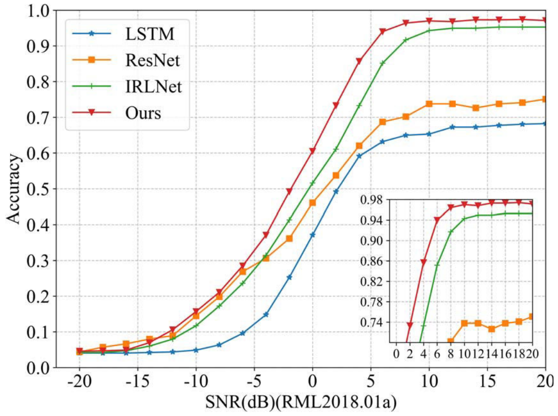
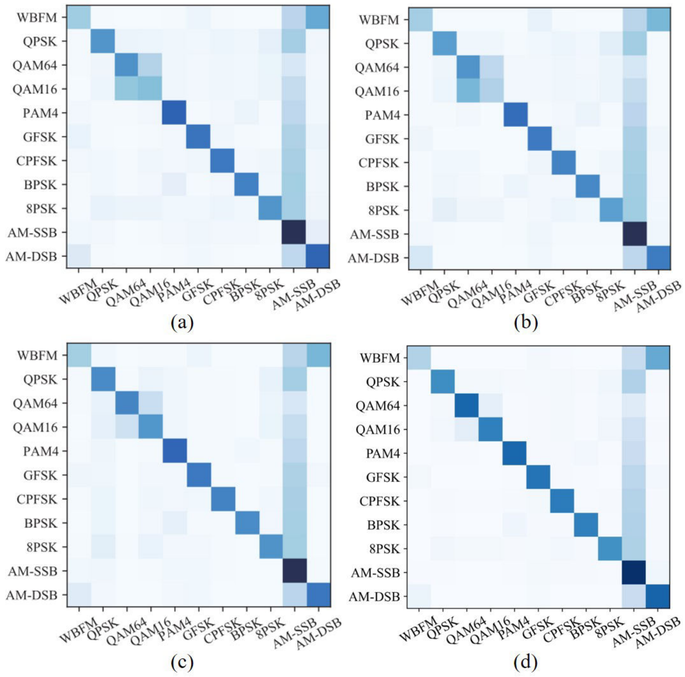


Fig. 5: Recognition accuracy on different SNR of SigFormer and other models on RML2018.01a.

As shown in Tab. I, on the RML2016.10a, RML2016.10b, RML2018.01a datasets, the SigFormer outperforms the state-of-the-art (SOTA) method by 1.41%, 0.73%, and 3.89% respectively on the average recognition accuracy, and outperforms the SOTA by 0.90%, 0.78% and 2.23% respectively on the highest recognition accuracy. More detailed, the SigFormer obtained higher recognition accuracy than other deep learning models under the vast majority of SNRs, as shown in Fig. 4 and Fig. 5.



(a) (b) (c) (d)

Fig. 6: Overall normalized confusion matrix of (a) 1D-ResNet (b) LSTM (c) SigNet (d) Ours on RML2016.10a.

D. Analysis on confusion matrix

A series of normalized confusion matrices of different methods on the entire testing set covering all SNRs are shown in Fig.6, which are calculated from all testing samples averagely and can reflect the overall recognition performance. It can be seen that the problem of confusion between QAM64 and QAM16 is more serious on 1D-ResNet and LSTM, the results on SigNet could alleviate this problem to a certain extent, while SigFormer can further release this confusion, since WBFM and AM-DSB are continuous modulations, the obvious features between them are very weak on complex panels, which causes all four models to confuse them easily [18], but SigFormer is less likely to misclassify AM-DSB as WBFM. On confusion matrices, it can be seen SigFormer obtained higher classifying probabilities on most modulation types.

E. Classification performance by modulation type

From the modulation type in Fig. 7, it can be seen that on the RML2016.10a, the recognition accuracy of AM-SSB has always remained at a high level, which means that AM-SSB has better noise resistance ability and can obtain higher average accuracy. When the SNR is no less than 0dB, the accuracy of all modulation types except WBFM exceeds or approaches 90% on both datasets, the main reason is that WBFM is easily misclassified by AM-DSB. On the RML2016.10b, the recognition performance of each modulation type is obvious different when the SNR is low, and the recognition accuracy of most modulation types is close to 100% when the SNR is above 0dB.

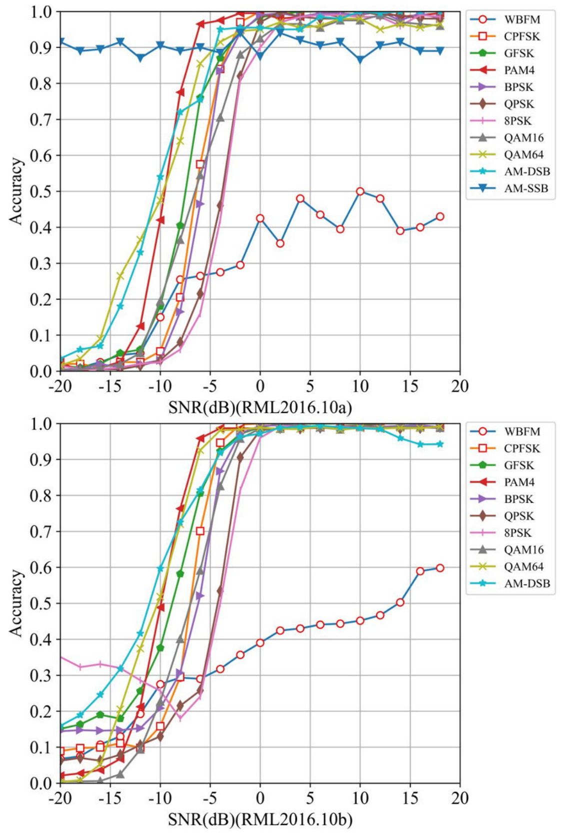


Fig. 7: Recognition performance of different modulation types.

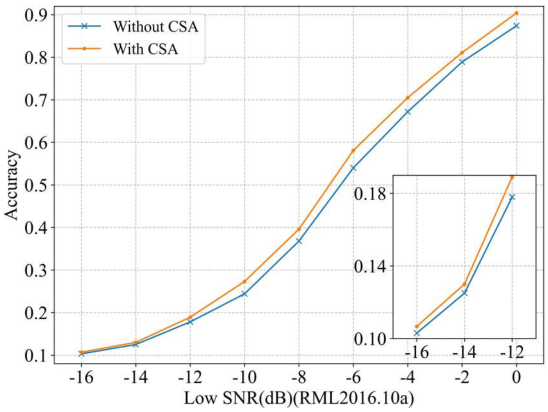


Fig. 8: The effect of CSA on low SNR.

TABLE II: Ablation study on model components

PYD	LSA	CSA	LFF	WarmUp	Accuracy↑	Params↓
					61.42%	0.143M
✓					61.55%	0.143M
✓	✓				61.88%	0.143M
✓	✓	✓			62.91%	0.148M
✓	✓	✓	✓		63.45%	0.158M
✓	✓	✓	✓	✓	**63.71%**	0.158M

TABLE III: Configuration with efficient versions.

Dataset	Models	Dimensions	Accuracy↑	Params↓
RML2016.10a	SigNet [15]	-	62.30%	23.52M
	Ours-Small	[4, 8, 16, 32]	63.69%	**0.044M**
	Ours-Large	[8, 16, 32, 64]	**63.71%**	0.158M
RML2016.10b	MCformer [13]	-	65.04%	0.073M
	Ours-Small	[4, 8, 16, 32]	65.29%	**0.044M**
	Ours-Large	[8, 16, 32, 64]	**65.77%**	0.158M
RML2018.01a	IRLNet [28]	-	60.07%	0.318M
	Ours-Small	[8, 16, 32, 64]	63.56%	**0.158M**
	Ours-Large	[16, 32, 64, 128]	**63.96%**	0.593M

F. Ablation study

An ablation study is performed on SigFormer with the non-pyramid structure model built by the original Transformer encoder module [9, 13] as the baseline to verify the effectiveness of the pyramid structure (PYD), warm up learning rate optimization strategy (WarmUp), and LSA, CSA, and LFF components. The experimental model is trained on the RML2016.10a training set and evaluated on the testing set. As shown in Tab. II, the gains of PYD, LSA, CSA, LFF, WarmUp on the average recognition accuracy of the model are 0.13%, 0.33%, 1.03%, 0.54%, 0.26%, respectively. It should be mentioned that the CSA performs better on lower SNRs samples. For instance, within the low SNR range of -16dB to 0dB, the average recognition accuracy with and without CSA is 45.51% and 43.26%, respectively, and the extra gain obtained by CSA is 2.25%, which exceeds the results on the wider SNR range (i.e., -20dB to 18dB) by 1.03%, as shown in Fig. 8. Moreover, within extremely lower SNR range (from -16dB to -12dB), the extra gains obtained by CSA are from 0.4% to 1.1%, which shows the great capability of SigFormer's modulation recognition on low SNR samples.

And in Tab. III, it can be seen that the SigFormer model can be scaled to small versions by more narrow channel configurations with fewer parameter counts, which is more efficient and achieved impressive recognition results on all datasets. When the SigFormer is configured with a smaller feature dimension at each stage, the average recognition accuracy also exceeds the baseline with extremely fewer parameters, which will be helpful for those scenarios demanding efficient models.

V. Conclusion

In this paper, we propose a SigFormer model for signal modulation recognition based on Transformer architecture, which is built on pyramid Transformer structure to extract high-dimension discriminative features for signal classification tasks. Experimental results on two benchmarks show that SigFormer achieved more excellent recognition performance than most previous deep learning models with fewer parameters, and it is more robust on low SNR signals.

References

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

[2] J. L. Xu, W. Su, and M. Zhou, "Likelihood-ratio approaches to automatic modulation classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, pp. 455–469, 2011.

[3] D.-C. Chang and P.-K. Shih, "Cumulants-based modulation classification technique in multipath fading channels," *IET Commun.*, vol. 9, pp. 828–835, 2015.

[4] S. Huang, Y. Yao, Z. Wei, Z. Feng, and P. Zhang, "Automatic modulation classification of overlapped sources using multiple cumulants," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 6089–6101, 2017.

[5] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," *ArXiv*, vol. abs/1602.04105, 2016.

[6] 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 Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.

[7] K. Liu, W. Gao, and Q. Huang, "Automatic modulationb recognition based on a dcn-bilstm network," *Sensors (Basel, Switzerland)*, vol. 21, 2021.

[8] Y. Wu, X. Li, and J. Fang, "A deep learning approach for modulation recognition via exploiting temporal correlations," *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, 2018.

[9] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[10] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *ArXiv*, vol. abs/2103.13413, 2021.

[11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *ArXiv*, vol. abs/2102.12122, 2021.

[12] H. Liu, X. Miao, C. Mertz, C. Xu, and H. Kong, "Crackformer: Transformer network for fine-grained crack detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3783–3792.

[13] S. Hamidi-Rad and S. Jain, "Mcformer: A transformer based deep neural network for automatic modulation classification," *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2021.

[14] T. J. O'Shea and N. E. West, "Radio machine learning dataset generation with gnu radio," 2016.

[15] Z. Chen, H. Cui, J. Xiang, K. Qiu, L. Huang, S. Zheng, S. Chen, Q. Xuan, and X. Yang, "Signet: A novel deep learning framework for radio signal classification," *IEEE Transactions on Cognitive Communications and Networking*, 2021.

[16] G. B. Tunze, T. Huynh-The, J.-M. Lee, and D.-S. Kim, "Sparsely connected cnn for efficient automatic modulation recognition," *IEEE Transactions on Vehicular Technology*, vol. 69, pp. 15 557–15 568, 2020.

[17] Z. Zhang, H. Luo, C. Wang, C. Gan, and Y. Xiang, "Automatic modulation classification using cnn-lstm based dual-stream structure," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 521–13 531, 2020.

[18] J. Xu, C. Luo, G. Parr, and Y. Luo, "A spatiotemporal multi-channel learning framework for automatic modulation recognition," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1629–1632, 2020.

[19] S. Wei, Q. Qu, X. Zeng, J. Liang, J. Shi, and X. Zhang, "Self-attention bi-lstm networks for radar signal modulation recognition," *IEEE Transactions on Microwave Theory and Techniques*, vol. 11, no. 69, pp. 5160 – 5172, September 2021.

[20] S. Lin, Y. Zeng, and Y. Gong, "Learning of time-frequency attention mechanism for automatic modulation recognition," *IEEE Wireless Communications Letters*, 2022.

[21] R. Müller, S. Kornblith, and G. E. Hinton, "When does label

smoothing help?” in *NeurIPS*, 2019.

[22] Z. Wang and T. Oates, “Imaging time-series to improve classification and imputation,” *ArXiv*, vol. abs/1506.00327, 2015.

[23] S. Peng, H. Jiang, H. Wang, H. Alwageed, Y. Zhou, M. M. Sebdani, and Y. dong Yao, “Modulation classification based on signal constellation diagrams and deep learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, pp. 718–727, 2019.

[24] E. Perenda, S. Rajendran, and S. Pollin, “Automatic modulation classification using parallel fusion of convolutional neural networks,” *IEEE WIRELESS COMMUNICATIONS*, 2019.

[25] N. E. West and T. J. O’Shea, “Deep architectures for modulation recognition,” *2017 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pp. 1–6, 2017.

[26] K. Tekbıyık, A. R. Ekti, A. Görçin, G. Karabulut-Kurt, and C. Keçeci, “Robust and fast automatic modulation classification with cnn under multipath fading channels,” *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1–6, 2020.

[27] D. Hong, Z. Zhang, and X. Xu, “Automatic modulation classification using recurrent neural networks,” *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pp. 695–700, 2017.

[28] H. Yang, L. Zhao, G. Yue, B. Ma, and W. Li, “Irlnet: A short-time and robust architecture for automatic modulation recognition,” *IEEE Access*, vol. 9, pp. 143 661–143 676, 2021.

[29] E. Blossom, “Gnu radio: tools for exploring the radio frequency spectrum,” *Linux Journal*, vol. 2004, p. 4, 2004.

[30] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2019.

[31] ——, “Sgdr: Stochastic gradient descent with warm restarts,” *arXiv preprint arXiv:1608.03983*, 2016.

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.