Cformer: A hybrid CNN-Transformer framework for 12-lead ECG denoising

Yewei Gan
School of Computer Science and Engineering
Sun Yat-Sen University
Guangzhou, China
ganyw3@mail2.sysu.edu.cn

Yanchong Xie
School of Computer Science and Engineering
Sun Yat-Sen University
Guangzhou, China
xieych37@mail2.sysu.edu.cn

Mingyue Cui*
School of Computer Science and Engineering
Sun Yat-Sen University
Guangzhou, China
*Corresponding author: cuimy@mail2.sysu.edu.cn

Kai Huang
School of Computer Science and Engineering
Sun Yat-Sen University
Guangzhou, China
huangk36@mail.sysu.edu.cn

***Abstract*—The electrocardiogram (ECG) signal serves as a crucial medical indicator, is prone to interference from various noise sources, leading to misdiagnosis of heart disease. However, previous denoising methods focus more on single-lead ECG, but ignore the correlation of each lead, which is not suitable for 12-lead ECG denoising. In this paper, we propose a novel hybrid CNN-Transformer framework named Cformer to explore the 12-lead ECG denoising method, which mainly includes a structured self-attention (SSA) module and a multi-scale feature aggregation (MSFA) module. For the SSA module, we use a two-stage attention strategy to capture the long-range dependencies and global features between different ECG leads according to their categories. For the MSFA module, we fuse local and global features through convolutions of different scales, which gives our model a wider field of view and produces better performance. Besides, we employ channel and spatial attention to make our model adaptively pay more attention to the critical ECG leads and time series. Experimental results show that our method achieves state-of-the-art performance compared with others on the 12-lead ECG benchmark PTB-XL dataset for four different types of noise.**

Keywords—12-lead ECG denoising, CNN, Transformer, biomedical signal processing, feature aggregation

I. Introduction

Electrocardiogram (ECG) signals play a crucial role in the diagnosis of heart disease (e.g., heartbeat classification [1]), and are considered the primary tool in clinical practice [2]. By analyzing the electrical activity of the heart, ECG signals provide valuable insights into the overall health of the cardiovascular system. Despite the critical role ECG signals play in medical diagnostics, they are frequently compromised by noise originating from both the patient's body and external factors. This susceptibility to interference can significantly contribute to the potential for misdiagnosis by medical professionals. The common ECG signal noise mainly includes the following categories: baseline wander (BW), electrode motion (EM), and muscle artifact (MA). Different types of noise have different characteristics. For example, BW noise is characterized by baseline offset and up and down fluctuations, which is mainly generated by unconscious autonomous movements of the human body; EM noise shows large amplitude and strong randomness, is caused by changes in the impedance and potential of the human skin; MA noise results from tremors caused by muscle contraction of the human body, and manifests as irregular and rapidly changing waveforms.

Due to the complex nature of the ECG signal and the presence of various types of noise, ECG denoising is a tough challenge. Fortunately, many scholars have conducted in-depth research on removing noise from ECG signals. [3] compares the performance of various Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filters for denoising ECG signals. [4] proposes IEMD-ATD method that combines empirical mode decomposition (EMD) and adaptive threshold to improve the decomposition quality and stability of raw ECG. Compared to traditional ECG denoising methods, network models demonstrate a more effective way of removing noise from ECG signals. [5] adopt a one-dimensional cyclic generative adversarial network to denoise artifact noise. With the generative neuron, this method avoids dependence on any prior distribution of ECG signals while enhancing the model's nonlinearity. [6] combines CNN and Transformer network to extract temporal information in ECG signal, and a new link constraint is introduced to the loss function to enhance the classification ability. These methods ignore the correlation of 12-lead ECG, making them only suitable for single-lead ECG denoising. Therefore, it is necessary to develop an effective 12-lead ECG denoising method to achieve high-quality and high-fidelity ECG signal reconstruction.

In this paper, we propose a novel hybrid CNN-Transformer architecture called Cformer for 12-lead ECG denoising, which combines the advantages of CNN for extracting local features and Transformer for extracting global features. Furthermore, the specifically designed structured self-attention module are used to explore the correlations between different ECG leads based on their categories. And the multi-scale feature aggregation module is designed to effectively fuse the local and

This work was supported by the Basic Research Operating Expenses of Universities-Young Teachers Cultivation Programs (No. 24qnpy140).

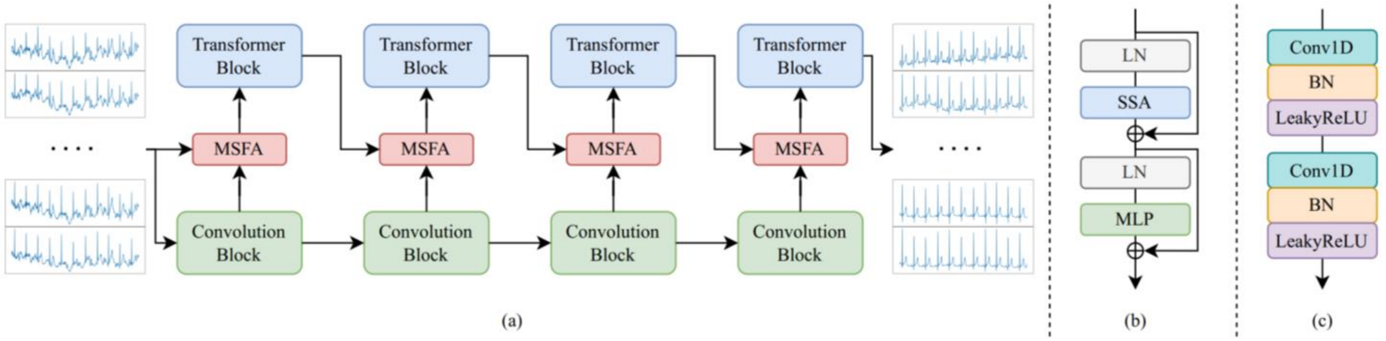


Fig. 1: The overall structure of proposed Cformer (a), the detailed transformer block (b), and convolution block (c).

global features. The contributions of our work can be summarized as follows:

- We propose a hybrid CNN-Transformer architecture to capture both local and global features of 12-lead ECG for better denoising performance.
- We design the structured self-attention module with a two-stage attention strategy to further explore the relationship between 12 different ECG leads based on different categories.
- To enhance the detailed representation ability of Transformer, we design the multi-scale feature aggregation module by fusing local features with global features on different scales.

II. Method

A. Parallel CNN-Transformer Architecture

CNN as one of the most common architectures, is widely used in the field of computer vision due to its effectiveness of capturing the local features and edge information. While Transformer demonstrates a strong ability to capture long-range dependencies and global features, and has shined in the field of natural language processing. To combine the advantages of the above two networks, we design the parallel CNN-Transformer architecture to explore both global and local enhanced features of 12-lead ECG for better denoising performance.

The detailed structure of our proposed method is shown in Fig. 1, which mainly consists of two parts, the CNN branch and the Transformer branch. Specifically, for the CNN branch, we use N consecutive convolution blocks to extract local features. At the same time, with the increase of network depth, the kernel size becomes larger and larger, which enlarges the receptive field and gives our model a better global vision in the deep layers of the network. Besides, we employ channel attention after each convolution operation to make our model adaptively focus on the critical ECG leads. For the Transformer branch, we first fuse the extracted local features and global features by our proposed MSFA module, then the fused features are fed into the transformer block to further capture the long-range dependencies. We also employ spatial attention followed by multi-layer perceptron layer to pay more attention to the key ECG time series.

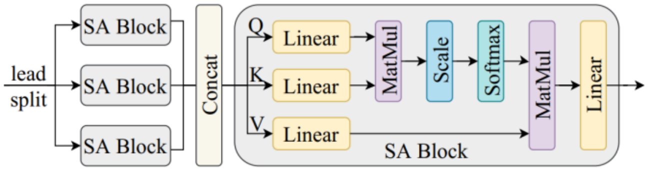


Fig. 2: Structured Self Attention (SSA).

B. Structured Self Attention

According to the placement of electrodes, the 12-lead ECG can be categorized into 3 different types, which are limb leads (leads I, II, and III), augmented limb leads (leads aVR, aVL, and aVF) and chest leads (leads V1 to V6). The leads in the same category have more similar correlations than others. Therefore, to further discover the relationships between different ECG leads with the same category, we design the structured self-attention (SSA) module to capture the long-range dependencies.

The detailed structure of SSA is illustrated in Fig. 2. Specifically, for the input 12-lead ECG, we first split it with respect to the above categories. Then, all 3 types of ECG leads are fed into the self-attention (SA) blocks to obtain the refined features of local types. After that, the local refined features are concatenated along the lead dimension and fed to the final SA block to obtain the fine-grained features of global types.

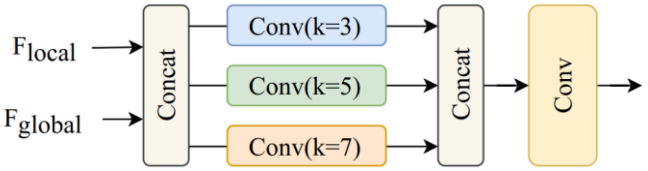


Fig. 3: Multi Scale Feature Aggregation (MSFA).

C. Multi Scale Feature Aggregation

To efficiently aggregate the global features extracted by Transformer and local features extracted by convolution, we design the multi-scale feature aggregation (MSFA) module to perform feature fusion operations at three different scales, which provides our model with multiple visions and results in better performance.

The data flow of our proposed MSFA is shown in Fig. 3. Specifically, we first concatenate the local and global features

Table. 1: The denoising results on different noise types and intensities.

Method	Noise Type	0 dB			5 dB		
		SNR_{out}↑	$RMSE$↓	PRD(%)↓	SNR_{out}↑	$RMSE$↓	PRD(%)↓
DNN [7]	BW	8.7876	0.1621	42.19	9.3832	0.1504	38.27
	EM	8.1631	0.1772	44.80	8.9016	0.1608	40.12
	MA	8.0035	0.1814	46.59	8.7337	0.1649	41.09
	Mix	8.2295	0.1750	45.08	9.0019	0.1585	39.81
FCN-DAE [7]	BW	11.458	0.1231	28.92	12.546	0.1093	25.93
	EM	7.6408	0.1846	46.33	9.3156	0.1559	37.30
	MA	8.6595	0.1659	41.35	10.386	0.1393	33.17
	Mix	8.1142	0.1758	43.79	9.8196	0.1479	35.19
SRD [8]	BW	24.061	0.0310	7.620	24.445	0.0296	7.327
	EM	24.504	0.0289	7.013	25.032	0.0273	6.682
	MA	24.170	0.0301	7.307	24.630	0.0285	7.002
	Mix	24.445	0.0291	7.084	24.933	0.0276	6.773
Cformer (**Ours**)	BW	27.236	0.0222	5.508	28.242	0.0195	4.840
	EM	27.723	0.0203	4.987	28.773	0.0179	4.455
	MA	27.491	0.0207	5.092	28.530	0.0183	4.548
	Mix	27.647	0.0205	5.064	28.732	0.0180	4.484

along the channel dimension. Then, we use 3 convolutions with different kernel sizes (3, 5, 7) to extract multi-scale features. After a concatenate operation, we use the final convolution layer to perform the feature aggregation to obtain the fused local and global features.

III. Experiments

A. Datasets

1) PTB-XL Dataset: The PTB-XL ECG dataset [9] is a large dataset of 21799 clinical 12-lead ECGs from 18869 patients of 10-second length, all sampled at 100 Hz or 500 Hz. We use the 100 Hz version as the ECG ground truth.

2) MIT-BIH Noise Stress Test Database: The MIT-BIH Noise Stress Test Database [10] includes 12 half-hour ECG recordings and 3 half-hour typical noises named BW, EM, and MA, all sampled at 360 Hz. For consistency, we resample them at 100 Hz. The Mix noise is generated by mixing the 3 types of noise in proportion.

B. Implementation Details

The depth N of our model is set to 4. To evaluate the robustness of our model under different noise intensities, we choose three different SNR levels (-1 dB, 3 dB, 7 dB) for training and another two levels (0 dB, 5 dB) for testing. The training pairs are generated by adding randomly sampled noises to the ECG ground truth. Similar to [7, 8], We use the MSE loss function and adopt the Adam optimizer to train the parameters of our model, the learning rate is set to 2×10^{-3} initially. We use an adaptive training strategy to reduce the learning rate automatically based on the loss metric. Specifically, if the loss does not decrease for 20 consecutive epochs, we reduce the learning rate to half of the previous value. The total training epochs are set to 300. The training process is carried out on an NVIDIA A100 server using PyTorch version 2.1 as the implementation language.

C. Comparison with Baseline Methods

We compare our method against methods DNN [7], FCN-DAE [7], and SRD [8]. We use the following three metrics to evaluate the denoising performance, i.e., SNR, RMSE, and PRD. The experimental results are shown in Table. 1. From the table, we can observe that our method outperforms all the other baseline methods on four types of noise with different noise intensities, as we expected. Compared with SRD, the output SNR of our method increases by 3.519 dB, and the average RMSE and PRD decrease by 0.0093 and 2.22 %, respectively. The experimental results demonstrate the effectiveness of our method.

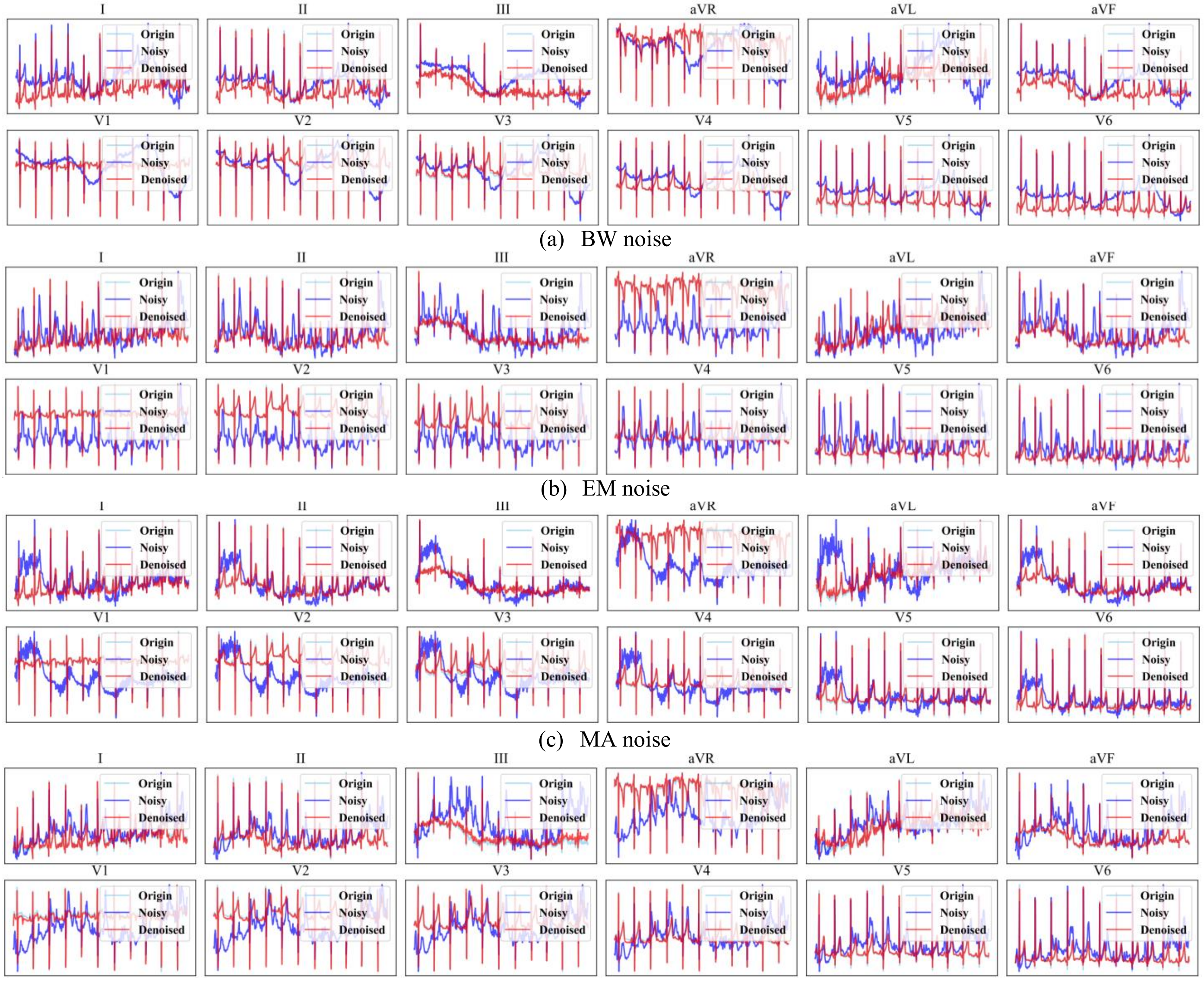
As shown in Fig. 4, we also visualize the denoising results of our method for 12 different ECG leads under different noise types. In the figure, the red part represents the denoised ECG signal, while the sky-blue and blue parts denote the ECG ground truth and the ECG contaminated by noise, respectively. We can observe that the red part almost completely covers the sky-blue part, which means that our method has achieved a similarity that is extremely close to the ground truth. The visualization results also demonstrate the great denoising performance of our method.

IV. Conclusion

In this paper, a hybrid CNN-Transformer model named Cformer is proposed for 12-lead ECG denoising. We explore an appropriate way to make full use of multi-lead ECG data through a two-stage attention strategy, which is beneficial for discovering the correlations between different ECG leads with the same category. Moreover, we fuse the extracted local and global features at different scales, which expands the model's field of view and further improves the denoising ability.

Acknowledgment

This work was supported by the Basic Research Operating Expenses of Universities-Young Teachers Cultivation Programs (No. 24qnpy140).



(a) BW noise

(b) EM noise

(c) MA noise

(d) Mix noise

Fig. 4: The visualization of our method with different noise types at 5 dB noise intensities, where ’Origin’ is the ground truth, ’Noisy’ and ’Denoised’ represent noisy ECG signals and our denoising results, respectively.

REFERENCES

[1] Wang Y, Zhou G, Yang C. Interpatient heartbeat classification using modified residual attention network with two-phase training and assistant decision[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 72: 1-15.

[2] E. Zvuloni, J. Read, A. H. Ribeiro, A. L. P. Ribeiro, and J. A. Behar, “On merging feature engineering and deep learning for diagnosis, risk prediction and age estimation based on the 12-lead ecg,” IEEE Transactions on Biomedical Engineering, 2023.

[3] N. Das and M. Chakraborty, “Performance analysis of fir and iir filters for ecg signal denoising based on snr,” in 2017 third international conference on research in computational intelligence and communication networks (ICRCICN). IEEE, 2017, pp. 90–97.

[4] M. Zhang and G. Wei, “An integrated emd adaptive threshold denoising method for reduction of noise in ecg,” Plos one, vol. 15, no. 7, p. e0235330, 2020.

[5] Kiranyaz S, Devecioglu O C, Ince T, et al. Blind ECG restoration by operational cycle-GANs[J]. IEEE Transactions on Biomedical Engineering, 2022, 69(12): 3572-3581.

[6] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, “Constrained transformer network for ecg signal processing and arrhythmia classification,” BMC Medical Informatics and Decision Making, vol. 21, no. 1, pp. 1–13, 2021.

[7] H.-T. Chiang, Y.-Y. Hsieh, S.-W. Fu, K.-H. Hung, Y. Tsao, and S.-Y. Chien, “Noise reduction in ecg signals using fully convolutional denoising autoencoders,” IEEE Access, vol. 7, pp. 60 806–60 813, 2019.

[8] Y. Hou, R. Liu, M. Shu, X. Xie, and C. Chen, “Deep neural network denoising model based on sparse representation algorithm for ecg signal,” IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1–11, 2023.

[9] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, “Ptb-xl, a large publicly available electrocardiography dataset,” Scientific data, vol. 7, no. 1, pp. 1–15, 2020.

[10] G. Moody, W. Muldrow, and R. Mark, “The mit-bih noise stress test database,” Computers in Cardiology, pp. 381–384, 1984.