

Dual-Stream Attention-TCN for EMG Removal from a Single-Channel EEG

Jun Lu, Ruihan Cai, Zhichao Guo, Qiyu Yang, Kan Xie*, and Shengli Xie, IEEE Fellow

Abstract—Long-term and mobile healthcare applications have increased the use of single-channel electroencephalogram (EEG) systems. However, electromyography (EMG) artifacts often disturb EEGs. The lack of spatial correlation, diversity of waveforms, and time-varying overlap make eliminating EMG interference from a single-channel EEG difficult. To overcome these challenges, we create DSATCN, a dual-stream learning model that makes use of multi-level and multi-scale temporal dependencies in different frequency bands to perform robust EEG reconstruction. The first DSATCN stream extracts low-frequency band EEG features with reduced EMG interference. The second stream selectively combines the high-level features of the first stream with its own low-level features to refine the EEG reconstruction across the entire frequency band, lowering the risk of overfitting. Both streams employ a novel attention-based temporal convolution network (ATCN) to adaptively separate the overlapping features of EEGs and EMGs. The ATCN has multiple stages to represent various temporal dependencies at different levels. Each stage consists of multi-scale dilated convolutions and fast Fourier transform modulations, which efficiently enrich the receptive fields and establish global self-attention mechanisms. The stages' outputs are merged by relaxed attentional feature fusion modules, which bridge semantic gaps between features at various levels. Extensive experimental results on three semi-simulated datasets containing 318,700 samples show that the proposed model significantly outperforms the existing methods in EEG reconstruction accuracy. And its computational cost meets the criteria for real-time processing. Our code is available at <https://github.com/BaenRH/DSATCN>.

Index Terms—single-channel EEG, EMG removal, temporal convolutional network

I. INTRODUCTION

THE electroencephalogram (EEG) contains a wealth of brain information that is significant for clinical diagnosis and neurological research. With the emergence of portable and wearable medical devices, the use of a restricted number of EEG channels, or even a single channel, has become increasingly popular in mobile and long-term applications, such as epilepsy detection [1], sleep staging [2], and the development of brain-computer interfaces [3]. However, electromyography (EMG) frequently interferes with EEGs [4], which can seriously degrade the quality of EEG data analysis.

This work was supported in part by the National Natural Science Foundation of China Grant (62073086, 62273106 and 62373114). Corresponding author: Kan Xie. All authors are with the Automation School, Guangdong University of Technology, Guangzhou, 510006, China (e-mail: kxie@gdut.edu.cn).

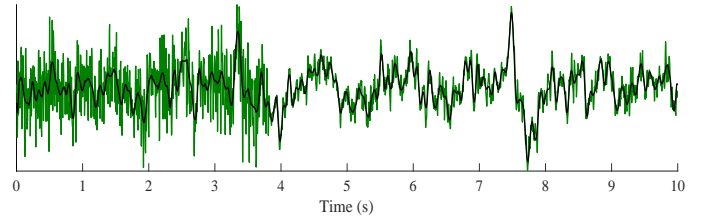


Fig. 1. The green line represents a single-channel EEG influenced by non-stationary EMG interference, whereas the black line represents the clean EEG reconstructed by our DSATCN.

Eliminating EMG interference from a single-channel EEG has long been regarded as a challenging task. This is due to the lack of spatial correlation across channels, the complexity of EEG waveforms, which are susceptible to EMGs with high amplitudes and a broad spectrum range, and the time-varying overlaps of EEG and EMG, as shown in Fig. 1.

Many approaches have been proposed in recent years to address this challenging task. They can be divided into two categories. The first group combines signal decompositions (e.g., ensemble empirical mode decomposition (EEMD) [1] and singular spectrum analysis (SSA) [2]) as well as blind source separation (BSS) (e.g., independent component analysis (ICA) [3], canonical correlation analysis (CCA) [4], and independent vector analysis (IVA) [5]) to exploit the correlations between latent components in a single channel, such as EEMD-ICA [6], EEMD-CCA [7], EEMD-IVA [8], and SSA-ICA [9]. These approaches first break up an EEG into multiple modes or signature signals, then apply BSS to extract EMG-free source components and reconstruct the clean EEG. Despite the utility of signal decompositions in addressing the nonstationarity and nonlinearity of EEGs [1] [2], their effectiveness is significantly influenced by manual parameter settings. These settings cannot well adapt to the intricate waveforms, as well as the time-varying contaminations. In addition, the subsequent BSSs usually require a visual inspection to select the source components, which is time-consuming and subjective.

The second group, which is based on deep learning (DL) techniques that use hierarchical nonlinear models, is more potent at representing complex EEG signals [10]–[14]. Their end-to-end inferences make source separation highly automated. For example, Sun et al. [10] introduced a multi-scale residual convolutional neural network (CNN) that can reconstruct EEGs more accurately than the ICA-based methods, the wavelet transforms, and the recursive least squares filter.

Zhang et al. [11] created a publicly available benchmark dataset to assess the single-channel EEG denoise methods. They found that the residual CNN [10] and the long short-term memory (LSTM) network [14] outperform the EMD method [15] and the bandpass filter for removing EMG. To improve the generalization performance, Zhang et al. [12] presented the novel CNN, which gradually increased feature dimensions and downsampled time sequences layer by layer. To make the separation of EEGs and artifacts more explicable, Yu et al. [13] proposed an encoder-decoder architecture with inception blocks. Although these DL approaches have shown promising results, their clinical application still has the following limitations. First, the model training is prone to overfitting [12]. The majority of the DL approaches employ general architecture designs while ignoring prior knowledge that EEG signals are less contaminated by EMG artifacts below 20 Hz [16]. Second, the EEG reconstruction is still inaccurate, especially during periods of high EMG contamination. The DL techniques mentioned above lack an effective mechanism for capturing various long-term temporal dependencies beyond the contamination periods.

In this paper, we design a dual-stream multi-task learning model named DSATCN to effectively leverage multi-level and multi-scale temporal dependencies for EMG removal. Both streams of DSATCN have an encoder-separator-decoder structure [17]. The first stream seeks to reconstruct EEG in the low-frequency band below 20 Hz, which is more robust to EMG interference. And the second stream combines the high-level outline features from the first stream with its own low-level detail feature to refine the EEG reconstruction across the entire frequency band. This multi-task learning paradigm provides multi-level supervision to guide the EEG feature extractions, improving the model generalization performance. In order to extract variable long-range temporal dependencies, we propose a novel attention-based temporal convolution network (ATCN) as the separator of each stream to adaptively filter out EEG components in the feature maps. The ATCN adopts a pyramid architecture that incorporates multi-scale, multi-level dilated convolutions (MMDCs), fast Fourier transform modulations (FFTM) and relaxed attentional feature fusion (RAFF) modules. Unlike the temporal convolution blocks with multiple dilated rates presented in the Conv-TasNet [17], the MMDCs have both multiple dilated rates and multiple kernel sizes, which significantly enrich the long-range receptive fields of the network. The FFTMs streamline the self-attention mechanism [18] by employing FFT, 1×1 convolutions, inverse FFT, and the Hadamard product. This approach allows the ATCN to efficiently acquire global context features, which is especially beneficial for processing EEG recordings with a high sampling rate. The concept of FFTM draws inspiration from two recent works, i.e., the Conv2Former [19] and the Res FFT-ReLU block [20]. Both of them prioritize the effectiveness of information extraction from extensive receptive fields. In contrast to the Conv2Former [19], our FFTM establishes connections between tokens using FFT, nonlinear activation, and inverse FFT rather than depth convolution, allowing us to take advantage of global information more efficiently during attention modulation. In comparison to the Res FFT-ReLU

block [20], our FFTM is more adaptable to the input features. The RAFF modules are responsible for fusing the features at various semantic levels. Because the optimal fusion could be outside the convex hull of the features, the complementary constraint on attention masks caused by sigmoid activation in the vanilla AFF [21] would restrict the model representation ability. Therefore, our RAFF omits the sigmoid activation and learns the attention masks for different level features, respectively. In addition, increasing the number of channels can greatly improve the model representation ability for diverse EEG temporal dependencies. However, by the end of the decoder, the fully connected layer after flattening will have a huge number of parameters (i.e., 67M, which is more than 90% of the total model parameters). Therefore, before the flattening layer, we utilize a pair of 1×1 convolutions in channel and temporal dimensions to reduce the size of a portion of the feature map. This partial flattening technique can largely reduce model size at the expense of a minor reduction in EEG reconstruction accuracy.

To the best of our knowledge, the proposed DSATCN is the first customized DL model that explicitly leverages the distribution discrepancy between EMG and EEG in the frequency domain. The main contributions of this paper are summarized as follows:

- A dual-stream multitask learning architecture is developed to adaptively combine the multi-level features from different frequency bands, enhancing EEG reconstruction, which reduces the overfitting risk.
- A novel attention-based TCN is designed to adaptively separate the EEG and EMG feature maps. Leveraging multi-scale and multi-level dilated convolutions, this model captures diverse local temporal dependencies while efficiently incorporating fast Fourier transform modulations to establish a robust self-attention mechanism.
- A relaxed attentional feature fusion module is proposed to more flexibly merge features from different semantic levels, and a partial flattening technique is proposed by the end of the decoder to efficiently reduce model size.
- Extensive experimental results on three semi-simulated datasets, totaling 318,700 samples, demonstrate that our DSATCN consistently surpasses current approaches in EEG reconstruction accuracy. Furthermore, a real-data experiment shows that DSATCN can effectively remove EMG and retain detailed EEG information.

II. DUAL-STREAM ATTENTION-BASED TEMPORAL CONVOLUTION NETWORK

In this section, we introduce our DSATCN. First, we show the notations and their descriptions in Table I. Then, we describe the network.

A. Network Architecture

Fig. 2 depicts the architecture of DSATCN. The first stream is responsible for low-frequency band EEG denoising. The second stream is responsible for full-frequency band EEG denoising. Both streams have an encoder-separator-decoder structure. A pair of single-layer 1-D convolutions are used

TABLE I
NOTATIONS AND ACRONYMS

| Notation | Description |
|---|--|
| <i>ATCN</i> | Attention-based temporal convolution network |
| <i>MMDC</i> | Multi-scale, multi-level dilated convolutions |
| <i>FFTM</i> | Fast Fourier transform modulation |
| <i>RAFF</i> | Relaxed attentional feature fusion |
| $\tilde{x}, x_l \in \mathbf{R}^T$ | Mixed EEG recordings in the whole frequency band and the low frequency band |
| $s, s_l \in \mathbf{R}^T$ | Ground true EEG signals in the whole frequency band and the low frequency band |
| $\tilde{s}, \tilde{s}_l \in \mathbf{R}^T$ | Reconstructed EEG signals in the whole frequency band, and the low frequency band |
| $D, D_l \in \mathbf{R}^{T_f \times C}$ | Low-level feature maps generated by the encoders |
| $D' \in \mathbf{R}^{T_f \times C}$ | Fused feature maps generated by the RAFF block |
| $M, M_l \in \mathbf{R}^{T_f \times C}$ | Masks generated by the separators |
| $Z, Z_l \in \mathbf{R}^{T_f \times C}$ | High-level feature maps of EEG components after separations |
| $X, O_1, O_2 \in \mathbf{R}^{T_f \times C}$ | The input and output feature maps of a 1-D dilated convolution block or a FFTM block |
| T, T_f, C | The numbers of time-sampling points, temporal tokens, and feature channels |
| \odot, \oplus, \otimes | Hadamard product, addition, and 1x1 convolution |

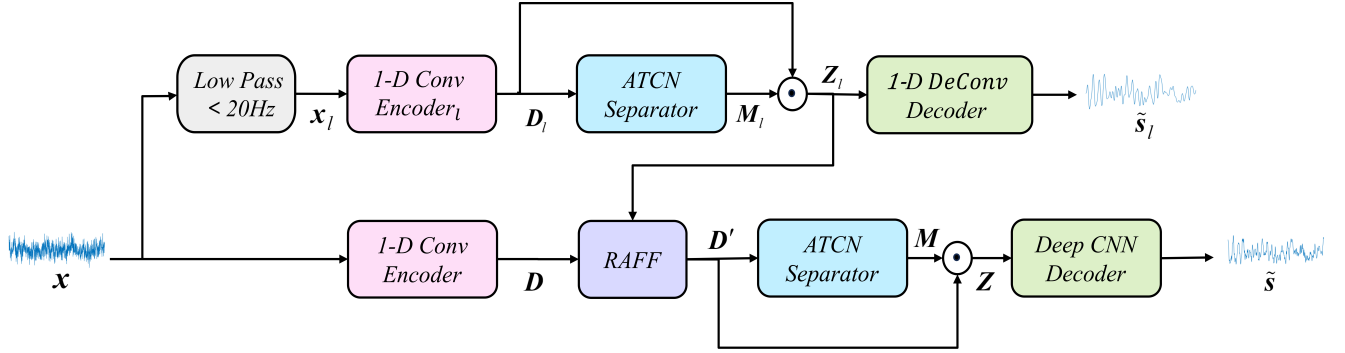


Fig. 2. Architecture overview of the DSATCN. x is the mixed EEG signal. \tilde{s}_l and \tilde{s} represent the reconstructed low-frequency band EEG and full-frequency band EEG, respectively.

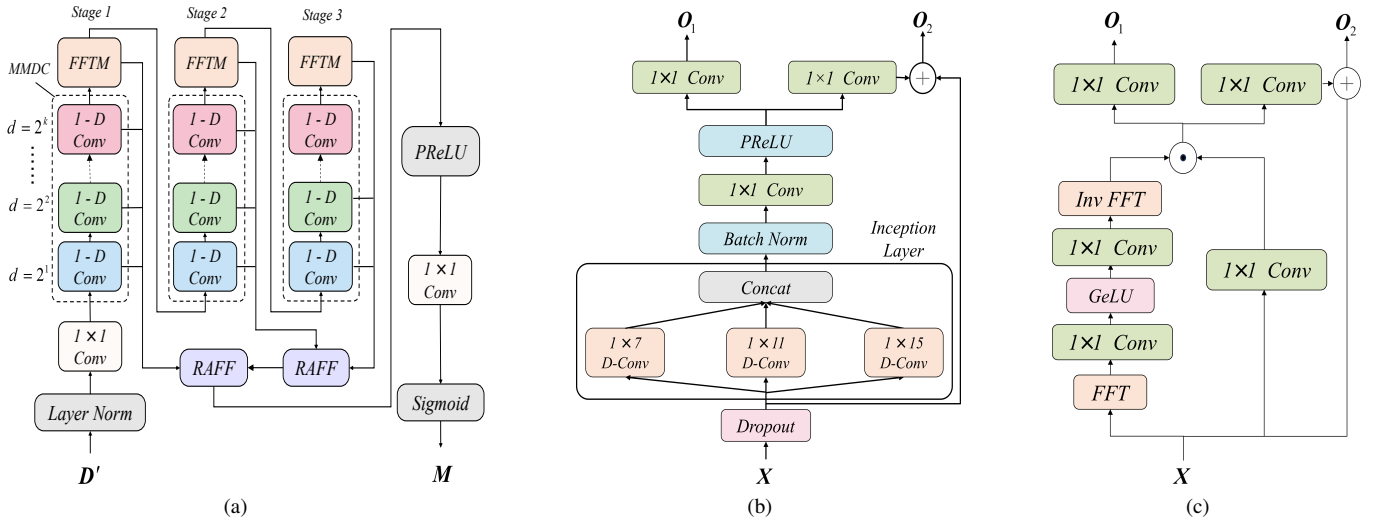


Fig. 3. Detailed architecture of the proposed ATCN separator (a), comprising 1-D dilated convolution blocks (b) and FFTM blocks (c). d denotes the dilation factor. Each block receives the input X from the previous block and outputs O_1 and O_2 to next block.

as the encoders to extract the low-level feature maps. The separators are composed of multi-stage ATCNs, which learn the masks to adaptively filter out the high-level feature maps of EEG components. The high-level feature maps are then sent into the decoders to reconstruct clean EEG waveforms. Because full-frequency EEG waveforms contain substantially more detailed information than low-frequency EEG waveforms, the first stream's decoder is a 1-D de-convolution layer, and the second stream's decoder is a deep CNN with a partial flatten layer. In particular, a RAFF module combines the first stream's high-level features with the second stream's low-level features to aid high-level feature learning in the second stream. The RAFF modules are also used to incorporate different stage information in ATCNs, enabling a more flexible integration of multi-level content. This design provides multi-level supervision information for network learning via a cross-branch gradient flow. Next, the components of DSATCN will be described in detail.

B. Attention-based temporal convolution network for separating the feature maps

Fig. 3(a) presents the pyramid architecture of the multi-stage ATCN module. It is an ensemble of multi-scale multi-level dilated convolutions (MMDCs), fast Fourier transform modulations (FFTM) and RAFFs. This design enables the separators to fuse multi-scale temporal dependencies at multi-level and multi-stages, thereby significantly improving the generalization performance.

1) *Multi-scale multi-level dilated convolution*: Fig. 3(b) shows the structure of the 1-D convolution blocks in a MMDC, which is similar to the Conv-Tasnet [17]. However, there are three modifications: i) Inspired by ConvNeXt [22], we perform depth-wise dilated convolution (D-Conv) first before point-wise convolution, which can reduce the floating-point operations (FLOPs). ii) The inception layer with varying length kernel sizes enriches the extraction of EEG features at different scales. iii) A channel-temporal wise dropout layer is deployed before each inception layer to improve the robustness.

2) *Fast Fourier transform modulation*: Although MMDCs greatly extend the receptive fields, they lack the adaptability to capture global information. To overcome this constraint, we embed FFTMs between MMDCs to generate multi-stage global features. Fig. 3(c) demonstrates the detailed structure of FFTM. Unlike the transformer [18], FFTM realizes the self-attention mechanism by modulating the feature map. It is implemented as follows:

i) Uses the FFT to efficiently convert the feature map into the frequency domain, based on the symmetric properties of the discrete Fourier transform (DFT) [23].

$$\mathbf{V} = \mathcal{F}(\mathbf{X}) \quad (1)$$

where $\mathbf{X} \in \mathbf{R}^{T_f \times C}$ is the input feature map, T_f and C are the numbers of temporal tokens and feature channels, function \mathcal{F} denotes FFT, $\mathbf{V} \in \mathbb{C}^{(0.5T_f+1) \times 2C}$ is composed of the real and imaginary parts of DFT coefficients that contain global temporal information, and \mathbb{C} denotes complex domain.

ii) Applies 1×1 convolution and a nonlinear activation to extract frequency domain features. As the analysis presented in

[20] shows, the nonlinear activation in the frequency domain enables the network to learn global contexts.

$$\mathbf{Y} = \mathcal{F}^{-1}[\sigma(\mathbf{V} \otimes \mathbf{W}_1) \otimes \mathbf{W}_2]$$

where \mathcal{F}^{-1} is the inverse FFT, σ is the Gaussian error linear unit (GELU) [24], $\mathbf{W}_1 \in \Omega^{(2C) \times d}$ and $\mathbf{W}_2 \in \Omega^{d \times (2C)}$ are parameters of 1×1 convolutions.

iii) Performs the Hadamard product between the outcomes of the inverse FFT and the features after channel interaction and subjects these results to individual processing using 1×1 convolutions. The generated feature maps $\mathbf{O}_1, \mathbf{O}_2 \in \mathbf{R}^{T_f \times C}$ are sent to the subsequent block and the skip connection path, respectively.

$$\begin{aligned} \mathbf{O}_1 &= (\mathbf{Y} \odot (\mathbf{X} \otimes \mathbf{W}_3)) \otimes \mathbf{W}_4 \\ \mathbf{O}_2 &= (\mathbf{Y} \odot (\mathbf{X} \otimes \mathbf{W}_3)) \otimes \mathbf{W}_5 + \mathbf{X} \end{aligned} \quad (2)$$

where $\mathbf{W}_{3,4,5} \in \mathbf{R}^{C \times C}$ are parameters of the 1×1 convolutions. The Hadamard product offers global temporal modulation. It enables dynamic feature fusion, which adaptively enhances the feature extracted by 1-D convolution blocks. Compared to the transformer [18], FFTM reduces the computation complexity of the self-attention mechanism from $O(4T_f \cdot C^2 + 2C \cdot T_f^2)$ to $O(4d \cdot C + C^2 + 2C \cdot T_f \cdot \log_2 T_f)$, which is more efficient for processing EEG recordings with a high sampling rate. The idea of FFTM draws inspiration from two recent works, i.e., the Conv2Former [19] and the Res FFT-ReLU block [20]. Both of them prioritize the effectiveness of information extraction from extensive receptive fields. In contrast to the Conv2Former [19], our FFTM establishes connections between tokens using FFT rather than depth convolution, allowing us to take advantage of global information more efficiently during attention modulation. In comparison to the Res FFT-ReLU block [20], our FFTM is more adaptable to the input features.

3) *Relaxed attentional feature fusion*: Since features from the two frequency band streams and different stages in an ATCN have different levels of semantic information and their contributions for EMG removal are non-stationary, simply aggregating these features via fixed weights, such as addition and concatenation, will degrade the generalization performance of our model. Therefore, we propose RAFF to adaptively generate the fusion weights based on the multi-scale feature contents. Our RAFF is a variant of vanilla AFF [21], which removes the sigmoid activation and learns the attention masks for different stream features, respectively. This modification eliminates the complementary constraint on attention masks. It enables simultaneously suppressing the noisy features from both streams and subtracting the common irrelevant components, thereby improving the flexibility of feature fusion. As shown in Fig. 4, RAFF consists of two branches to aggregate the local and global context information along the channel dimension. Given the feature maps \mathbf{F}_1 and $\mathbf{F}_2 \in \mathbf{R}^{C \times T}$ from the two streams, the local channel context \mathbf{L}_c is computed via a bottleneck structure as follows:

$$\mathbf{L}_c = \text{ReLU}(\mathcal{B}((\mathbf{F}_1 + \mathbf{F}_2) \otimes \mathbf{W}_l)) \quad (3)$$

where $\mathbf{W}_l \in \mathbf{R}^{C \times C/r}$ represents the 1×1 convolution parameter matrix, r symbolizes the channel reduction ratio, \mathcal{B}

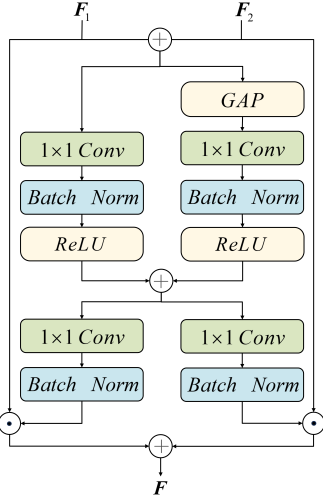


Fig. 4. Structure of the relaxed RAFF.

denotes batch normalization, while ReLU implies the rectified linear unit. L_c maintains the nuanced aspects of the lower-level features. And the global channel context $G_c \in \mathbf{R}^C$ is formulated as follows:

$$G_c = \text{ReLU}(\mathcal{B}(\text{GAP}(F_1 + F_2) \otimes W_g)) \quad (4)$$

where GAP denotes the global average pooling, $W_g \in \mathbf{R}^{C \times C/r}$ is the 1×1 convolution parameter matrix. Then L_c and G_c are merged as:

$$P = L_c \oplus G_c \quad (5)$$

where \oplus denotes the broadcasting addition. We use $P \in \mathbf{R}^{C \times T}$ to separately generate the attention masks for different stream features, and fuse the masked components as follows:

$$F = F_1 \odot \mathcal{B}(P \otimes W'_1) + F_2 \odot \mathcal{B}(P \otimes W'_2) \quad (6)$$

where W'_1 and $W'_2 \in \mathbf{R}^{C/r \times C}$ are the 1×1 convolution parameter matrices, $F \in \mathbf{R}^{C \times T}$ denotes the fusion output. By deploying the RAFF modules, our DSATCN can adaptively integrate the inconsistent semantic information from multiple stages and levels.

C. Deep CNN and partial flattening for decoding EEG waveforms

Fig. 5 depicts the structure of our deep decoder, which includes depthwise separable convolution blocks, a partial flatten layer, and a fully connected layer. With the stacking of convolution layers and nonlinear activations, we steadily increase the number of feature channels and down sample the time sequences. This strategy has been demonstrated to be effective in boosting generalization performance [12]. It could be related to the extraction of diverse features that are not sensitive to temporal positions. The flatten and fully connected layers reconstruct the EEG waveforms with features across channels and time intervals. However, given a great number of channels and a long EEG sequence, e.g., 512 channels, 128 tokens, and 1024 EEG time samples, the fully connected layer will have massive parameters, $512 \times 128 \times 1024 \approx 67M$,

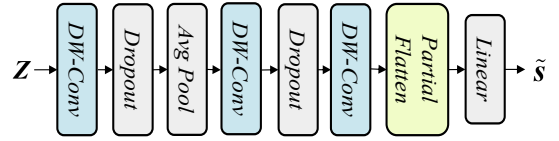


Fig. 5. Structure of the full-band decoder.

which is more than 94% of the total model parameters. In order to mitigate redundant features and reduce computational complexity, we propose a partial flattening approach for EEG waveform reconstruction. Given a feature map $U \in \mathbf{R}^{T_s \times C_l}$, the partial flattening is implemented as follows:

i) Segments U into two parts along the channel dimension as $U_1 \in \mathbf{R}^{T_s \times [p \cdot C_l]}$, $U_2 \in \mathbf{R}^{T_s \times [(1-p) \cdot C_l]}$, where $p \in (0, 1)$ is the proportion;

ii) Flattens U_1 to a vector $q_1 \in \mathbf{R}^{p \cdot C_l T_s}$;

iii) Linearly projects U_2 to lower dimensions as $U'_2 = A \times U_2 \times B \in \mathbf{R}^{m \times n}$, where $A \in \mathbf{R}^{m \times T_s}$, $B \in \mathbf{R}^{[(1-p) \cdot C_l] \times n}$, $m, n \ll \min\{T_s, C_l\}$, then flattens U'_2 to a vector $q_2 \in \mathbf{R}^{mn}$;

iv) Concatenates q_1 and q_2 to a feature vector $q \in \mathbf{R}^{m \cdot n + p \cdot C_l T_s}$.

Thus, given small p , m , and n , the partial flattening approach can significantly reduce the feature dimensions and save computation for the following fully connected layer.

D. Loss Function

The DSATCN aims to simultaneously reconstruct the low-frequency band and full-frequency band EEG signals as close as possible to the ground truths. We use the relative absolute errors to define the loss function as

$$\text{loss} = \|s - \tilde{s}\|_1 / (\|s\|_1 + \epsilon) + \|s_l - \tilde{s}_l\|_1 / (\|s_l\|_1 + \epsilon) \quad (7)$$

where \tilde{s}_l and s_l are the reconstructed EEG and ground truth in the low-frequency band, $\|s_l - \tilde{s}_l\|_1 / (\|s_l\|_1 + \epsilon)$ performs as a regularization term demanding parts of feature channels in the full-band stream can accurately generate s_l . ϵ is set to 1×10^{-8} to prevent division by zero. The relative error makes the loss function well-suited for processing EEG signals with varying magnitudes. The \mathcal{L}_1 norm is used to lessen the effect of reconstruction errors caused by outlier samples in EEG recordings.

III. EXPERIMENTS

A. Experimental Paradigm

1) **Datasets:** We use three semi-simulated datasets and a real-life dataset recorded by different devices with various sampling rates and reference types to evaluate the effectiveness of DSATCN, which are described as follows:

Semi-simulated Datasets I: EEGdenoiseNet [11]. It includes 4514 clean EEG segments and 5598 EMG segments from 52 subjects. Each segment is sampled at 512 Hz with a duration of 2 seconds and pre-processed using the common average reference method. The EEG segments are expanded to match the EMG sample size through random sampling. Both EEG and EMG datasets are then divided into training, validation, and test sets with 4,478, 560, and 560 segments,

TABLE II
CONFIGURATIONS OF THE PROPOSED DSATCN

| | Low-frequency band stream | Full-band stream |
|-----------|--|--|
| Encoder | [Conv, 1×4 , 256, stride 4] | [Conv, 1×4 , 256, stride 4] |
| Separator | $\left[\begin{array}{l} \text{[Conv, } 1 \times 1, 64, \text{ stride 1]} \\ \left[\begin{array}{l} \text{DConv, } 1 \times 7, 64, \text{ stride 1} \\ \text{DConv, } 1 \times 11, 64, \text{ stride 1} \\ \text{DConv, } 1 \times 15, 64, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 256, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 64, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 64, \text{ stride 1} \end{array} \right] \times 6 \\ \left[\begin{array}{l} \text{Linear, } 128 \times 256 \\ \text{Linear, } 256 \times 128 \\ \text{Linear, } 128 \times 128 \end{array} \right] \times 1 \\ \left[\begin{array}{l} \text{Conv, } 1 \times 1, 16, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 64, \text{ stride 1} \end{array} \right] \times 2 \end{array} \right] \times s$ | $\left[\begin{array}{l} \text{[Conv, } 1 \times 1, 64, \text{ stride 1]} \\ \left[\begin{array}{l} \text{DConv, } 1 \times 7, 64, \text{ stride 1} \\ \text{DConv, } 1 \times 11, 64, \text{ stride 1} \\ \text{DConv, } 1 \times 15, 64, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 256, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 64, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 64, \text{ stride 1} \end{array} \right] \times 6 \\ \left[\begin{array}{l} \text{Linear, } 128 \times 256 \\ \text{Linear, } 256 \times 128 \\ \text{Linear, } 128 \times 128 \end{array} \right] \times 1 \\ \left[\begin{array}{l} \text{Conv, } 1 \times 1, 16, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 64, \text{ stride 1} \end{array} \right] \times 2 \end{array} \right] \times s$ |
| | [Conv, 1×1 , 256, stride 1] | [Conv, 1×1 , 256, stride 1] |
| RAFF | $\left[\begin{array}{l} \text{Conv, } 1 \times 1, 128, \text{ stride 1} \\ \text{Conv, } 1 \times 1, 256, \text{ stride 1} \end{array} \right] \times 2$ | |
| Decoder | [Deconv, 1×1 , 1, stride 1] | $\begin{array}{l} \text{[DWConv, } 1 \times 1, 512, \text{ stride 1]} \\ \text{[Avg Pooling, } 1 \times 2, \text{ stride 2]} \\ \text{[DWConv, } 1 \times 7, 512, \text{ stride 1]} \\ \text{[DWConv, } 1 \times 7, 512, \text{ stride 1]} \\ \text{[Fc, } 32832 \times 1024 \text{]} \end{array}$ |

respectively. According to [11], an EMG is randomly selected to mix with a clean EEG at the SNR_{RMS} within the range from -7 to 2 dB at the interval of 1 dB. The SNR_{RMS} is defined as

$$SNR_{RMS} = 10 \log_{10}(\|x_{EEG}\|_2 / \|\lambda x_{EMG}\|_2) \quad (8)$$

where x_{EEG} and x_{EMG} are the pure EEG and EMG segments, λ is a parameter adjusting the SNR_{RMS} of the mixed EEG. The mixed EEG is generated as $x = x_{EEG} + \lambda x_{EMG}$.

Semi-simulated Datasets II: MPIHCBS-CAP. It is a combination of the Max Planck Institute for Human Cognitive and Brain Sciences (MPIHCBS) dataset [25] and the Cyclic Alternating Pattern (CAP) sleep database [26]. The MPIHCBS dataset comprises 54 clean EEG recordings acquired from 27 individuals. Each recording has a duration of 54 to 84 seconds and was captured using nine electrodes with a sampling rate of 200 Hz. The EEG electrodes placement followed the 10–20 International System with a specific referencing approach: odd-indexed electrodes (FP1, F3, C3, etc.) were referenced to the left mastoid, even-indexed electrodes (FP2, F4, C4, etc.) to the right mastoid, and central electrodes (Fz, Cz, Pz) were referenced to the average of the left and right mastoids. The CAP database consists of 16 two-channel EMG recordings acquired from 108 individuals. Each recording has a duration of 6.8 to 30 hours and was caught using different sampling rates ranging from 128 to 512 Hz. In order to facilitate the synthesis of the semi-simulated dataset, the EEG and EMG recordings are resampled to 256 Hz and segmented every 4 seconds. The EEGs from the MPIHCBS dataset are partitioned into training, validation, and test sets, consisting of 5016, 627, and 627 segments, respectively. An equivalent number of EMG segments are randomly selected from the CAP database and mixed with the EEG segments in the same way as EEGdenoiseNet [11].

Semi-simulated Datasets III: ISRUC-III [27]. It comprises 8589 recordings from 10 subjects. Each recording is 30 seconds long, sampled at 200 Hz, and includes data from six EEG channels and two EMG channels. The EEG channels use A1 and A2 references, placed in the left and right earlobes, respectively. The recordings are cut into 309,204 5-second segments. Unlike semi-simulated datasets I and II, the EEG and EMG segments used to generate the mixed signals are synchronized. To ensure data validity and reliability, we adopted the individual-based division strategy to build the training, validation, and test sets and conduct 5-fold cross-validation.

Real-life Dataset: BioSource EEG [28]. It was recorded by a long-term epilepsy monitoring unit using an average reference montage. As shown in Fig. 6, the 10-second EEG data are sampled at 250 Hz with 21 channels. Severe EMG artifacts exist in the EEG recording between 0–3.9 seconds on channels F7, T3, T5, C3 and T1, and between 5–10 seconds on channels F8, T4, C4 and P4.

2) Evaluation Metrics: Our evaluation metrics includes SNR , relative root mean square errors in the temporal domain ($RRMSE_t$) and spectral domain ($RRMSE_s$), and correlation coefficient (CC), which are computed as

$$SNR = 10 \log_{10}(\|s\|_2^2 / \|s - \tilde{s}\|_2^2) \quad (9)$$

$$RRMSE_t = \|\tilde{s} - s\|_2 / \|s\|_2 \quad (10)$$

$$RRMSE_s = \|PSD(\tilde{s}) - PSD(s)\|_2 / \|PSD(s)\|_2 \quad (11)$$

$$CC = Cov(\tilde{s}, s) / \sqrt{Var(\tilde{s})Var(s)} \quad (12)$$

where the SNR denotes the energy ratio between the clean EEG and the estimation error, the $PSD(\cdot)$ denotes the signal's power spectral density, $RRMSE_t$ and $RRMSE_s$ measure the deviations of an estimated EEG from the true value in

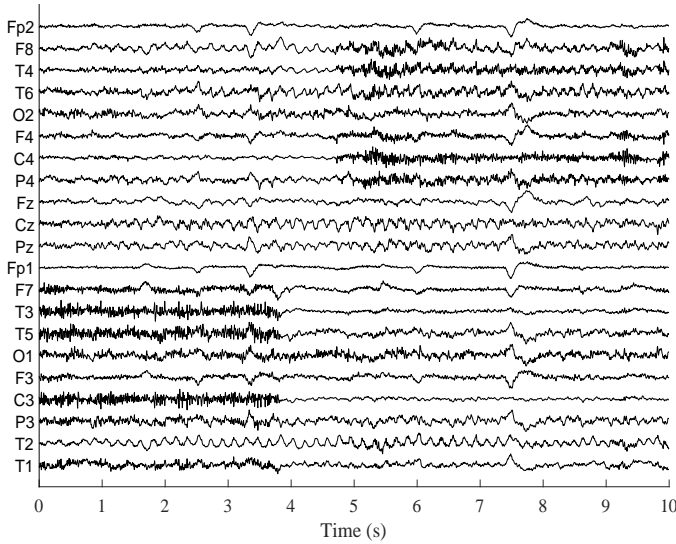


Fig. 6. The real-life EEG contaminated with EMG artifacts [28].

the temporal and frequency domains, and CC measures the similarity between the estimation and ground true EEGs.

3) *Implementation Details*: The configurations of DSATCN are listed in TABLE II. s represents the number of stage in Separator. In each 1-D dilated convolution block, the dropout probability is set to 0.1. As for the partial-flatten, the p , m and n are set to 0.5, 8 and 8. AdamW [29] is used to optimize the parameters of DSATCN. The initial learning rate is set to 0.001. We use a cosine learning rate decay schedule with warmup over 10% of training. For the EEGdenoiseNet and MPIHCBS-CAP datasets, we configure the training with 100 epochs and a batch size of 64. For the ISRUC-III dataset, the training is set to 20 epochs with a batch size of 16. The network model is implemented in Python 3.8 with Pytorch 2.0, training on a workstation with one Nvidia RTX 4090 GPU.

4) *Baseline Methods*: We compare DSATCN to the SOTA methods, including EEMD-ICA [6], SSA-ICA [9], 1-D ResCNN [10], Novel CNN [12] and DeepSeparator [13]. The BSS methods are implemented by the Matlab toolbox 'ReMAE' [30]: for EEMD-ICA, the noise level, the numbers of ensembles, and the intrinsic mode functions are respectively set to 0.2, 20 and $\lceil \log_2(n) \rceil$ (n is the number of sample points); for SSA-ICA, the window length and the number of decomposition components are set to 30 and 10, respectively, and the components with autocorrelation values ≤ 0.95 are removed as artifacts. The DL methods are implemented in Python 3.7 with Pytorch 1.11: for 1-D ResCNN, the network parameters are adjusted to accommodate the input segment length of the semi-simulated data; for Novel CNN and DeepSeparator, all parameters are set as same as those presented in [12] and [13].

B. Results and Discussions

1) *Results of Semi-Simulated Data*: Fig. 7 depicts the training and validation losses across epochs for three semi-simulated datasets, respectively. As shown, the training losses steadily decreased and converged after approximately 90 epochs for the EEGdenoiseNet and MPIHCBS-CAP datasets

TABLE III
COMPARISON OF AVERAGED EEG RECONSTRUCTION ACCURACY OBTAINED BY DIFFERENT METHODS.

| Dataset | Models | $RRMSE_s$ | $RRMSE_t$ | CC | SNR (dB) |
|----------------|---------------|---------------|---------------|---------------|----------------|
| EEG-DenoiseNet | EEMD-ICA | 2.4924 | 0.9892 | 0.6963 | 1.3211 |
| | SSA-ICA | 2.1909 | 0.9558 | 0.6981 | 1.8111 |
| | 1-D ResCNN | 0.6329 | 0.6272 | 0.7774 | 4.4702 |
| | DeepSeparator | <i>0.4611</i> | 0.6304 | 0.7814 | 4.7404 |
| | Novel CNN | 0.4685 | <i>0.4496</i> | <i>0.8637</i> | 8.2046 |
| | DSATCN | 0.3726 | 0.3530 | 0.9023 | 10.9798 |
| MPIHCBS-CAP | EEMD-ICA | 1.5292 | 0.8143 | 0.7075 | 2.9805 |
| | SSA-ICA | 1.1658 | 0.8150 | 0.7391 | 3.1487 |
| | 1-D ResCNN | 0.5977 | 0.6227 | 0.7975 | 4.6719 |
| | DeepSeparator | <i>0.4744</i> | 0.5476 | 0.8319 | 5.9633 |
| | Novel CNN | 0.4823 | <i>0.5248</i> | <i>0.8411</i> | 6.2686 |
| | DSATCN | 0.3542 | 0.4156 | 0.8892 | 8.7056 |
| ISRUC-III | EEMD-ICA | 0.3121 | 0.4864 | 0.8479 | 6.8529 |
| | SSA-ICA | 0.4623 | 0.5839 | 0.8324 | 5.6755 |
| | 1-D ResCNN | 0.2059 | 0.3690 | 0.9220 | 9.1747 |
| | DeepSeparator | <i>0.1626</i> | <i>0.3298</i> | <i>0.9341</i> | <i>10.3172</i> |
| | Novel CNN | 0.1747 | 0.3324 | 0.9326 | 10.2282 |
| | DSATCN | 0.1419 | 0.3043 | 0.9440 | 11.1398 |

* The best accuracies are in bold; the second-best are in italics.

and around 18 epochs for the ISRUC-III dataset. With the decrease in training losses, the validation losses gradually went down until the number of epochs reached around 85 epochs for the EEGdenoiseNet and MPIHCBS-CAP datasets and around 12 epochs for the ISRUC-III dataset. This result indicates the effectiveness of parameter learning for the proposed DSATCN.

TABLE III lists the EEG reconstruction accuracies averaged across the testing samples on the three semi-simulated datasets, respectively. Despite variations in acquisition devices, sample rates, and reference types among these datasets, the proposed DSATCN significantly outperformed all the baseline methods in $RRMSE_t$, $RRMSE_s$, CC , and SNR (paired t-test, $p < 10^{-4}$). Besides, Fig. 8 shows that DSATCN was also remarkably superior to the baseline methods in the medians of $RRMSE_t$, $RRMSE_s$, CC , and SNR (Wilcoxon signed rank test, $p < 10^{-4}$). The reconstruction accuracies of EEMD-ICA and SSA-ICA have large confidence intervals, which indicates that these BSS methods are tenuous to the variable EEG recordings. The 1D-ResCNN has branches at multiple scales. Each branch stacks the residual convolutional blocks. This architecture allows the model to represent variable EEG waveforms and significantly outperform ICA methods in all the accuracy metrics on the three datasets (Wilcoxon signed rank test, $p < 10^{-4}$). And the confidence intervals of 1D-ResCNN were much smaller than those of ICA methods. DeepSeparator also has multiple scale branches with residual convolutional blocks but is equipped with an attention mechanism to highlight the latent EEG components, which can adapt to the influences of non-stationary EMGs. The performance of DeepSeparator was better than that of 1D-ResCNN in most accuracy metrics. The novel CNN progressively downsampled EEGs and increased the number of channels layer by layer, which enlarged the receptive field and enriched the representation patterns, resulting in comparable performance with DeepSeparator, although without an attention mechanism. However, the multiple average pooling operations of Novel

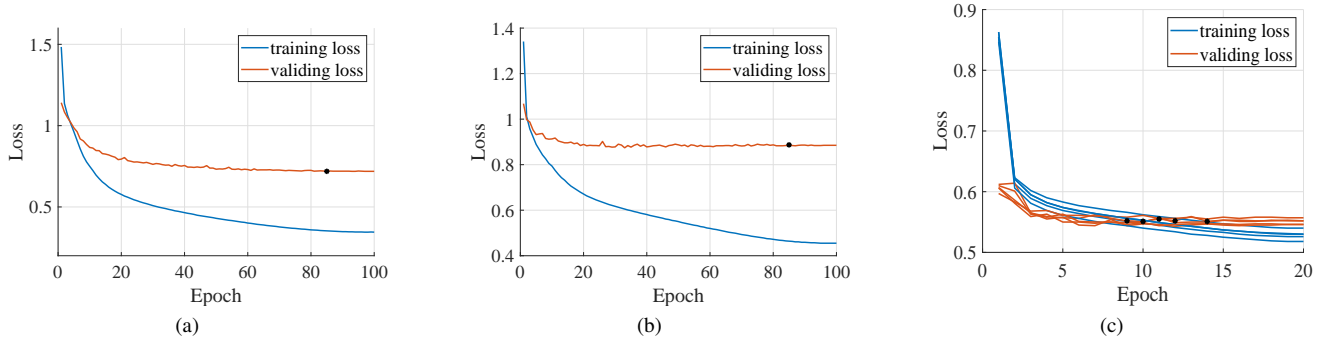


Fig. 7. Training and validation loss curves for the datasets EEGDenoiseNet (a), MPIHCBS-CAP (b), and ISRUC-III (c), respectively. The dark marks represent the lowest places in the validation loss curves.

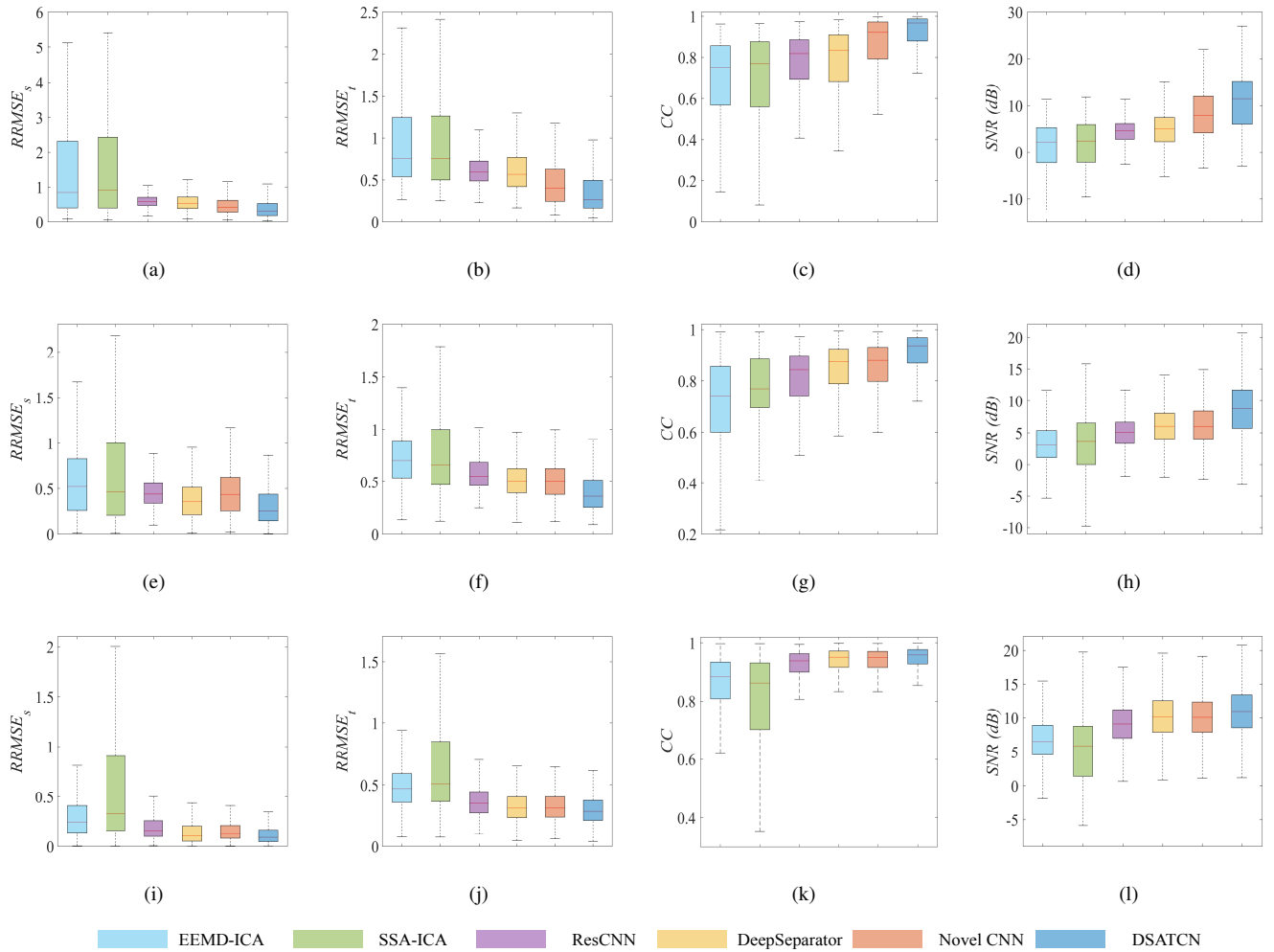


Fig. 8. The boxplots of EEG reconstruction accuracy obtained by different methods for the datasets EEGDenoiseNet (a-d), MPIHCBS-CAP (e-h), and ISRUC-III (i-l), respectively.

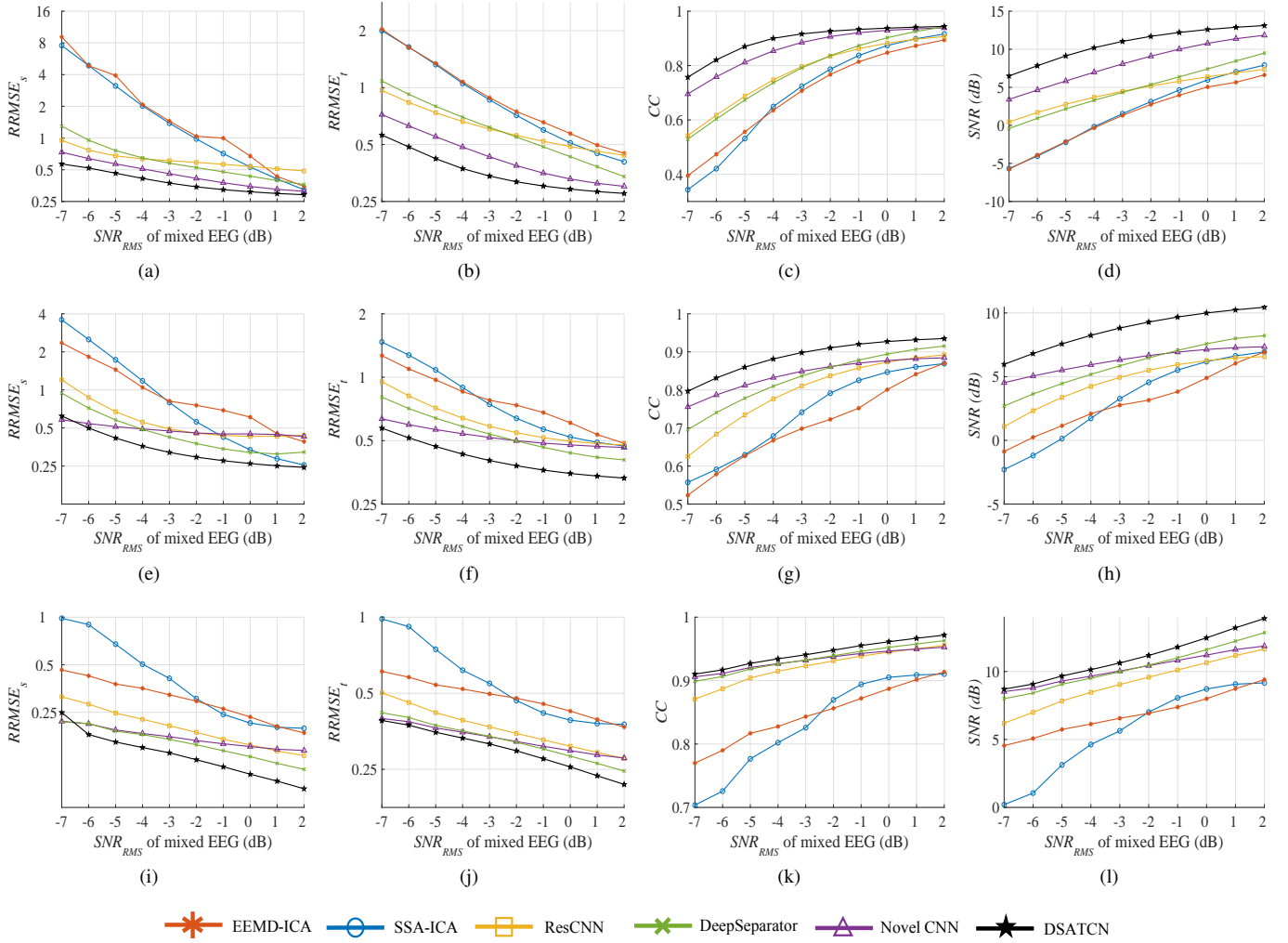


Fig. 9. Average EEG reconstruction accuracies of different methods across various SNR_{RMS} levels on datasets EEGDenoiseNet (a - d), MPIHCBS-CAP (c - h) and ISRUC-III (i - l).

CNN may degrade the resolution of long-range temporal dependencies. Neither 1D-ResCNN, DeepSeparator, nor Novel CNN can i) efficiently capture the global temporal information of high sample rate EEGs; ii) adaptively integrate features from different semantic levels; and iii) effectively use the prior knowledge that low-frequency band EEGs are less influenced by EMGs. To address these limitations, the proposed DSATCN incorporates the novel FFTM, RAFF, and dual stream architecture, which may contribute to its advantages in EEG reconstruction accuracy.

Fig. 9 demonstrates that DSATCN consistently achieved the best or second-best accuracy for EEG reconstruction across different SNR_{RMS} levels in all the evaluation metrics. The accuracies of EEMD-ICA and SSA-ICA quickly dropped when the SNR_{RMS} of mixed EEGs went down. This could be because of strong EMG inferences and the lack of reference information from other channels, which can cause statistical biases. Since the low-frequency band of EEGs is less contaminated by EMGs, the progressive downsampling strategy improved the noise resistance of Novel CNN at lower SNRs, particularly at -7 dB, where its performance was close to that of our DSATCN. Thanks to the attention

mechanism, DeepSeparator consistently outperformed 1D-ResCNN on the datasets MPIHCBS-CAP and ISRUC-III, maintaining commendable performance across different levels of noise contamination. However, on the EEGDenoiseNet dataset, DeepSeparator did not hold superiority to 1D-ResCNN in all the SNR_{RMS} levels. This could be owing to the higher sample rate (512 Hz) of the EEGDenoiseNet dataset, which limited the receptive field of convolutions and hence the effect of the local attention mechanism. While our DSATCN uses the RAFF to dynamically extract resilient EEG features from the low-frequency band to assist in the denoising of the full-band EEG, it also incorporates the FFTM to capture global temporal dependencies. This gave DSATCN an edge across a wide range of SNR_{RMS} levels on datasets with different EEG sample rates.

Table IV demonstrates that: i) The configurations with two-tuple and three-tuple kernel sizes outperform the single kernel size configurations in most of the accuracy metrics across the three datasets. This discovery aligns with the information presented in references [10] and [13], indicating that multi-scale convolutional kernels are more effective at modeling intricate EEG waveforms; ii) Using the same kernel size

TABLE IV

COMPARISON OF AVERAGED EEG RECONSTRUCTION ACCURACY WITH DIFFERENT KERNEL SIZES

| Dataset | Kernel sizes | RRMSE _s | RRMSE _t | CC | SNR (dB) |
|----------------|--------------|--------------------|--------------------|---------------|----------------|
| EEG-DenoiseNet | 7 | 0.3851 | 0.3628 | 0.9023 | 10.5501 |
| | 11 | 0.3826 | 0.3607 | 0.9036 | 10.5626 |
| | 15 | 0.3815 | 0.3593 | <i>0.9028</i> | 10.6420 |
| | 19 | 0.3859 | 0.3577 | 0.9026 | 10.7918 |
| | 7, 11 | 0.3810 | 0.3571 | 0.9012 | 10.8562 |
| | 11, 15 | 0.3815 | 0.3573 | 0.9017 | 10.8855 |
| | 7, 19 | 0.3797 | 0.3546 | 0.9026 | 10.9306 |
| | 7, 11, 15 | 0.3726 | 0.3530 | 0.9023 | 10.9798 |
| | 11, 15, 19 | <i>0.3775</i> | <i>0.3537</i> | 0.9016 | <i>10.9626</i> |
| MPIHCBS-CAP | 7 | 0.3648 | 0.4282 | 0.8854 | 8.3792 |
| | 11 | 0.3564 | 0.4258 | 0.8860 | 8.3936 |
| | 15 | 0.3713 | 0.4249 | 0.8873 | 8.4188 |
| | 19 | 0.3538 | 0.4225 | 0.8879 | 8.4464 |
| | 7, 11 | 0.3572 | 0.4206 | 0.8883 | 8.5205 |
| | 11, 15 | 0.3653 | <i>0.4191</i> | <i>0.8887</i> | <i>8.6214</i> |
| | 7, 19 | 0.3676 | 0.4207 | 0.8885 | 8.5571 |
| | 7, 11, 15 | <i>0.3542</i> | 0.4156 | 0.8892 | 8.7056 |
| | 11, 15, 19 | 0.3613 | 0.4202 | 0.8872 | 8.5822 |
| ISRUC-III | 7 | 0.1475 | 0.3057 | 0.9426 | 11.0328 |
| | 11 | 0.1430 | 0.3049 | 0.9429 | 11.0610 |
| | 15 | 0.1372 | 0.3047 | 0.9430 | 11.0737 |
| | 19 | 0.1470 | 0.3049 | 0.9429 | 11.0624 |
| | 7, 11 | 0.1389 | <i>0.3041</i> | 0.9431 | 11.0843 |
| | 11, 15 | <i>0.1366</i> | 0.3042 | 0.9433 | 11.0790 |
| | 7, 19 | 0.1386 | 0.3039 | <i>0.9434</i> | <i>11.0884</i> |
| | 7, 11, 15 | 0.1419 | 0.3043 | 0.9440 | 11.1398 |
| | 11, 15, 19 | 0.1359 | 0.3042 | 0.9432 | 11.0839 |

* The best accuracies are in bold; the second-best are in italics.

configuration [7, 11, 15] yielded the best or second-best results in most accuracy metrics on different datasets. Increasing the kernel sizes to [11, 15, 19] could negatively impact the accuracy. It implies that using oversize kernels in the proposed DSATCN may lead to a redundancy in the model's capability and hurt the generalization performance. Table V presents the EEG reconstruction accuracy with varying numbers of stages in ATCN. The three-stage configuration demonstrated superior performance across all accuracy metrics on the EEG-DenoiseNet and MPIHCBS-CAP datasets. Upon analyzing the ISRUC-III dataset with a sampling rate of 200 Hz, it was seen that the performance of the two-stage approach surpassed that of the three-stage approach in terms of $RRMSE_t$ and SNR . Furthermore, it exhibited a modest advantage in the $RRMSE_s$, which is a metric used for spectrum reconstruction. This finding suggests that cutting down the number of stages in ATCN could lower the risk of overfitting while still leaving enough model capacity to get rid of EMGs from EEG recordings that have a lower sampling rate.

2) *Results of Real-Life Data*: Since the real-life dataset has no ground truth, the temporal waveforms and spectral contents of reconstructed EEGs are adopted to qualitatively evaluate the EMG removal performance. As shown in Fig. 6, visible electrooculogram (EOG) artifacts exist in the real-life recordings. Thus we eliminate the EOGs first by using the Savitzky-Golay (SG) [31] filter. The SG filter is an efficient adaptive smoother, which can well eliminate the EOG and movement artifacts in EEG recordings. For BSS methods, the 10-s real-life data without EOG is input into the toolbox

TABLE V

COMPARISON OF AVERAGED EEG RECONSTRUCTION ACCURACY WITH DIFFERENT NUMBER OF STAGES IN ATCN

| Dataset | # Stage | RRMSE _s | RRMSE _t | CC | SNR (dB) |
|----------------|---------|--------------------|--------------------|---------------|----------------|
| EEG-DenoiseNet | 1 | 0.4178 | 0.3837 | 0.8940 | 10.004 |
| | 2 | 0.3965 | 0.3632 | <i>0.9009</i> | 10.6625 |
| | 3 | 0.3726 | 0.3530 | 0.9023 | 10.9798 |
| | 4 | <i>0.3738</i> | <i>0.3549</i> | 0.9023 | <i>10.8403</i> |
| MPIHCBS-CAP | 1 | <i>0.3593</i> | 0.4340 | 0.8827 | 8.1719 |
| | 2 | 0.3615 | 0.4256 | 0.8861 | 8.4133 |
| | 3 | 0.3542 | 0.4156 | 0.8892 | 8.7056 |
| | 4 | 0.3698 | <i>0.4186</i> | <i>0.8877</i> | <i>8.7032</i> |
| ISRUC-III | 1 | <i>0.1406</i> | 0.3077 | 0.9411 | 10.9835 |
| | 2 | 0.1366 | 0.3022 | 0.9440 | 11.1414 |
| | 3 | 0.1419 | <i>0.3043</i> | 0.9440 | <i>11.1398</i> |
| | 4 | 0.1409 | 0.3049 | <i>0.9430</i> | 11.0431 |

* The best accuracies are in bold; the second-best are in italics.

channel by channel. For DL methods, the network models are trained and validated on the semi-simulated dataset. To obtain better generalization performance on the real-life data, we resample the semi-simulated data to the sample rate of real-life data and extend the SNR_{RMS} levels of mixed EEGs from -7 to 15 dB. Then the real-life data without EOG are cut into 2-s sliding windows with strides of 100 sampling points and input into the network models in succession. The estimated EEGs of multiple sliding windows in overlapping parts are averaged to improve the reconstruction accuracy. Finally, the averaged EEGs are concatenated and added with the artifacts eliminated by the SG filter to evaluate the EMG removal performance.

Fig. 10 shows that SSA-ICA, EEMD-ICA and DSATCN can remarkably reduce the EMGs between 0–3.9 seconds in channel C3 and 5–10 s in channels F8, T4 and C4. However, DeepSeparator does not function well for these heavily contaminated EEGs. It may be because the inception-CNN modules without downsampling and flatten layers prevent DeepSeparator from capturing global channel-temporal dependencies of EEG signals. The reconstructed EEGs of SSA-ICA and EEMD-ICA are too smooth between 0–2 seconds in channel C3 and 3–4 seconds in channel T1, while the proposed DSATCN remains more details than BSS methods.

Fig. 11 displays that: i) the PSD curves of all the methods decrease notably in the higher frequency bands above 30 Hz. However, the PSD curves of the DeepSeparator follow those of the original EEGs at longer distances. Since EMGs are much stronger than EEGs in the higher frequency bands [32], DSATCN, EEMD-ICA and SSA-ICA are more potent than DeepSeparator in suppressing the EMGs; ii) in channels Fp1, F7 and F3, the PSD curves of SSA-ICA and EEMD-ICA keep decreasing within 10–30 Hz, while those of DSATCN have a rise around 20 Hz. Because the β rhythms of EEGs are mainly located within 14–30 Hz [32], SSA-ICA and EEMD-ICA may lead to the loss of β components, while DSATCN can preserve more valuable information. These results coincide with the observations in Fig. 10.

3) *Ablation Study*: To investigate the validity of the specific modifications in DSATCN, we conduct ablation experiments with the three semi-simulated datasets. Table VI reveals that

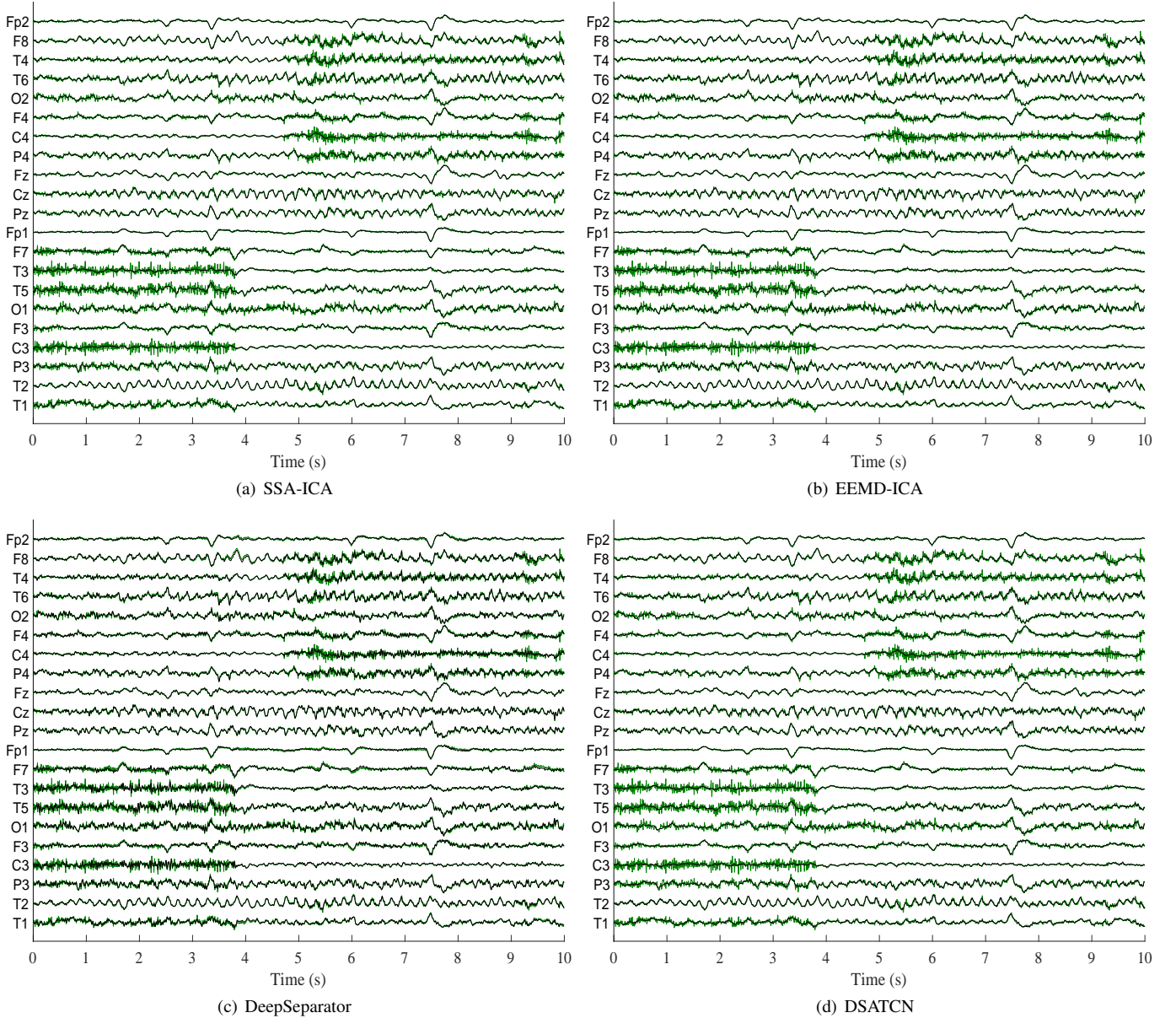


Fig. 10. EMG removal results of different methods on the real-life dataset. The green lines and black lines respectively denote the EEG waveforms before and after eliminating EMGs.

the dual-stream architecture markedly outperforms the single-stream model in terms of average $RRMSE_t$ and SNR (paired t-test, $p < 0.01$).

The results presented in Table VII indicate that the proposed RAFF method effectively integrates information from various hierarchical levels, resulting in significantly better performance compared to the addition [17] and concatenation methods with fixed fusion weights. This improvement is observed in terms of average values of $RRMSE_t$, CC , and SNR (paired t-test, $p < 0.01$). RAFF, in contrast to vanilla AFF, removes the constraints imposed by sigmoid activation and instead trains attention masks for different stream features. This leads to additional enhancements in average SNR of EEG reconstruction.

Table VIII shows that the both integrating the FFT-ReLU

TABLE VI
COMPARISON OF AVERAGED EEG RECONSTRUCTION ACCURACY WITH DIFFERENT ATCN ARCHITECTURE

| Dataset | Architectures | $RRMSE_s$ | $RRMSE_t$ | CC | SNR (dB) |
|----------------|---------------|---------------|---------------|---------------|----------------|
| EEG-DenoiseNet | Single-stream | 0.4180 | 0.3707 | 0.9039 | 10.1758 |
| | Dual-stream | 0.3726 | 0.3530 | 0.9023 | 10.9798 |
| MPIHCBS-CAP | Single-stream | 0.3540 | 0.4275 | 0.8857 | 8.3158 |
| | Dual-stream | 0.3542 | 0.4156 | 0.8892 | 8.7056 |
| ISRUC-III | Single-stream | 0.1487 | 0.3101 | 0.9414 | 10.8685 |
| | Dual-stream | 0.1419 | 0.3043 | 0.9440 | 11.1398 |

[20] and FFTM in each stage of ATCN can ameliorate the EEG reconstruction accuracies in different degrees, suggesting that global feature extraction of long-term dependencies is

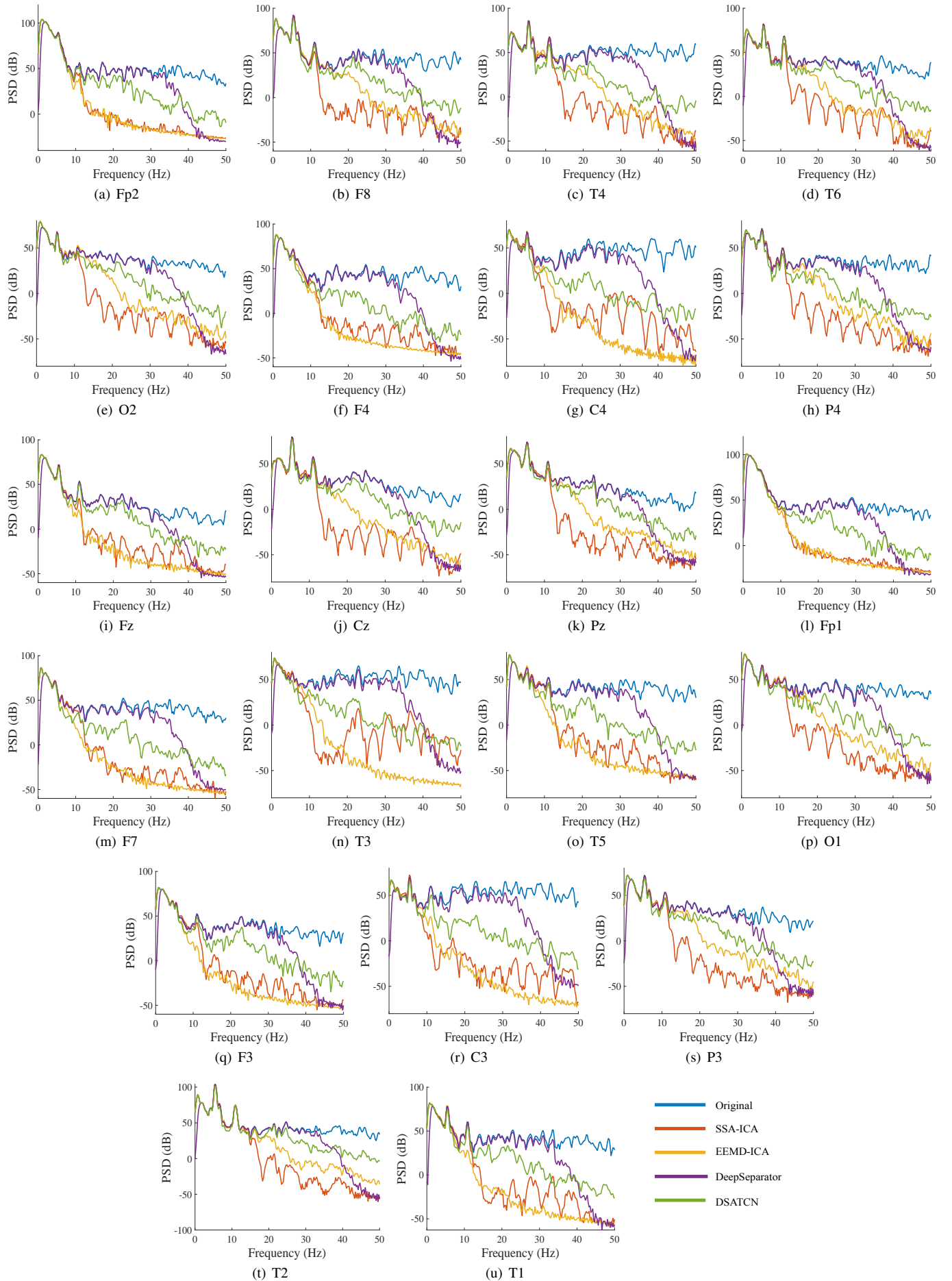


Fig. 11. Welch's PSDs of the EEG signals reconstructed by different methods in single-channel cases.

TABLE VII

EFFECT OF THE PROPOSED RAFFS DEPLOYED BETWEEN MULTIPLE STAGES AND STREAMS OF DSATCN

| Dataset | Fusion types | RRMSE _s | RRMSE _t | CC | SNR (dB) |
|----------------|------------------|--------------------|--------------------|---------------|----------------|
| EEG-DenoiseNet | Addition [17] | 0.4072 | 0.3704 | 0.8976 | 10.4467 |
| | Concatenation | 0.4155 | 0.3778 | 0.8876 | 10.5036 |
| | Vanilla AFF [21] | 0.3756 | 0.3544 | 0.9022 | 10.8727 |
| | Proposed RAFF | 0.3726 | 0.3530 | 0.9023 | 10.9798 |
| MPIHCBS-CAP | Addition [17] | 0.3579 | 0.4286 | 0.8840 | 8.3799 |
| | Concatenation | 0.3698 | 0.4225 | 0.8870 | 8.5288 |
| | Vanilla AFF [21] | 0.3582 | 0.4182 | 0.8904 | 8.5456 |
| | Proposed RAFF | 0.3542 | 0.4156 | 0.8892 | 8.7056 |
| ISRUC-III | Addition [17] | 0.1387 | 0.3114 | 0.9407 | 10.8532 |
| | Concatenation | 0.1344 | 0.3066 | 0.9425 | 11.0096 |
| | Vanilla AFF [21] | 0.1356 | 0.3043 | 0.9432 | 11.0773 |
| | Proposed RAFF | 0.1419 | 0.3043 | 0.9440 | 11.1398 |

TABLE VIII

EFFECT OF THE FFTM DEPLOYED IN EACH STAGE OF ATCNs

| Dataset | FFT types | RRMSE _s | RRMSE _t | CC | SNR (dB) |
|----------------|---------------|--------------------|--------------------|---------------|----------------|
| EEG-DenoiseNet | Without FFTM | 0.3828 | 0.3617 | 0.8993 | 10.6735 |
| | FFT-ReLU [20] | 0.3823 | 0.3592 | 0.8993 | 10.7490 |
| | Proposed FFTM | 0.3726 | 0.3530 | 0.9023 | 10.9798 |
| MPIHCBS-CAP | Without FFTM | 0.3480 | 0.4287 | 0.8834 | 8.3406 |
| | FFT-ReLU [20] | 0.3552 | 0.4199 | 0.8873 | 8.5911 |
| | Proposed FFTM | 0.3542 | 0.4156 | 0.8892 | 8.7056 |
| ISRUC-III | Without FFTM | 0.1372 | 0.3049 | 0.9431 | 11.0592 |
| | FFT-ReLU [20] | 0.1423 | 0.3049 | 0.9431 | 11.0611 |
| | Proposed FFTM | 0.1419 | 0.3043 | 0.9440 | 11.1398 |

TABLE IX

EFFECT OF THE PROPOSED PARTIAL FLATTENING DEPLOYED BY THE END OF THE DEEP CNN DECODER IN THE SECOND STREAM

| Dataset | Flatten layers | RRMSE _s | RRMSE _t | CC | SNR (dB) |
|----------------|--------------------|--------------------|--------------------|---------------|----------------|
| EEG-DenoiseNet | Full flattening | 0.3714 | 0.3519 | 0.9033 | 11.0253 |
| | Partial flattening | 0.3726 | 0.3530 | 0.9023 | 10.9798 |
| MPIHCBS-CAP | Full flattening | 0.3679 | 0.4156 | 0.8915 | 8.6859 |
| | Partial flattening | 0.3542 | 0.4156 | 0.8892 | 8.7056 |
| ISRUC-III | Full flattening | 0.1435 | 0.3049 | 0.9429 | 11.0324 |
| | Partial flattening | 0.1419 | 0.3043 | 0.9440 | 11.1398 |

propitious for EEG denoising. By introducing the modulation mechanism, the proposed FFTM becomes more adaptable to the input features than FFT-ReLU, and significantly improves the average $RRMSE_t$ and SNR (paired t-test, $p < 0.05$).

According to the results presented in Table IX, the impact of partial flattening is comparable to or slightly superior to that of full flattening in all three datasets. And as shown in Table X, employing partial flattening instead of full flattening can result in a reduction of over 46% of the parameters of DSATCN.

4) *Computational Efficiency Analysis*: In Table XI, we compare the parameter count and floating-point operations of our network when the stage number is set to 3 with those of other methods. It shows that DeepSeparator is the most efficient DL method with the fewest parameters and FLOPs. At the same time, the computational loads of Novel CNN and DSATCN are remarkably higher. The reasons are i) both Novel CNN

TABLE X

COMPARING THE COMPUTATION EFFICIENCIES OF DSATCN USING DIFFERENT FLATTEN LAYERS

| Flatten Manners | # Parameters (M) | FLOPs (M) |
|--------------------|------------------|----------------|
| Full flattening | 71.51 | 1062.40 |
| Partial flattening | 38.02 | 1029.19 |

TABLE XI

COMPUTATION EFFICIENCY COMPARISON OF DIFFERENT DL METHODS

| Models | # Parameters (M) | FLOPs (M) |
|--------------------|------------------|--------------|
| 1-D ResCNN [10] | 33.45 | 101.60 |
| Novel CNN [12] | 58.72 | 631.34 |
| DeepSeparator [13] | 0.30 | 31.91 |
| Proposed DSATCN | 38.02 | 1029.19 |

and DSATCN have much wider and deeper decoders with the flatten and fully-connected layers; and ii) the dual-stream architecture of DSATCN doubles the computation of the multi-stage ATCN containing inception layers, FFTM and RAFF modules. However, when running on a laptop equipped with an Intel(R) Core(TM) i5-6300HQ CPU @2.30GHz, DSATCN takes only 0.13 ± 0.01 second to process a 2-second EEG recording with a 512 Hz sampling rate and can work in real-time. For EEG recordings that do not exhibit a duration divisible by 2 seconds, an overlapping trick can be used to extract the segments. Subsequently, the denoising outcomes of the overlapped segments are averaged, allowing for more robust and accurate EEG denoising.

IV. CONCLUSIONS

In this paper, we propose a dual-stream attention-based temporal convolution network (DSATCN) for eliminating EMG from a single-channel EEG. This dual-stream design enables DSATCN to extract high-level EEG features from the low-frequency band as a reference to enhance EEG reconstruction in the full-frequency band, reducing the overfitting risk. The ATCN combines multi-scale multi-level dilated convolutions (MMDCs), fast Fourier transform modulations (FFTM) and relaxed attentional feature fusion (RAFF) modules to adaptively separate the EEG and EMG feature maps with varying overlaps in each stream. Especially, MMDCs effectively model a wide range of local temporal dependencies with multiple dilated rates and kernel sizes, while FFTMs efficiently construct a self-attention mechanism to globally modulate EEG features generated by MMDCs on multiple levels, significantly enhance the model robustness. The RAFF modules flexibly integrate the inconsistent semantic information from multiple stages and levels, remarkably improving the EEG reconstruction accuracies. We conducted extensive experiments on both the semi-simulated and real-world datasets. The results show that our DSATCN remarkably beats SOTA approaches in terms of EEG reconstruction accuracy. We also carried out ablation studies to demonstrate the contribution of each module described in this paper.

Nonetheless, our method has the following flaws: i) The model is quite large and has a longer inference time, despite

the fact that we used the partial flattening technique to reduce the model size by 46%; ii) Interference in EEG recordings comprises EOGs, electrocardiograms, and measuring noises in addition to EMGs. This can make EEG reconstruction more complicated. In the future, we will look into the re-parameterization techniques [33], [34] to improve inference efficiency. Furthermore, we intend to apply our approach to eliminate other types of noise in clinical EEG data.

REFERENCES

- [1] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.
- [2] R. Vautard, P. Yiou, and M. Ghil, "Singular-spectrum analysis: A toolkit for short, noisy chaotic signals," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1, pp. 95–126, 1992.
- [3] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.
- [4] W. De Clercq, A. Vergult, B. Vanrumste, W. Van Paesschen, and S. Van Huffel, "Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2583–2587, 2006.
- [5] X. Chen, H. Peng, F. Yu, and K. Wang, "Independent vector analysis applied to remove muscle artifacts in EEG data," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 7, pp. 1770–1779, 2017.
- [6] B. Mijović, M. De Vos, I. Gligorićević, J. Taelman, and S. Van Huffel, "Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 9, pp. 2188–2196, 2010.
- [7] K. T. Sweeney, S. F. McLoone, and T. E. Ward, "The use of ensemble empirical mode decomposition with canonical correlation analysis as a novel artifact removal technique," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 97–105, 2012.
- [8] X. Chen, A. Liu, H. Peng, and R. K. Ward, "A preliminary study of muscular artifact cancellation in single-channel EEG," *Sensors*, vol. 14, no. 10, pp. 18 370–18 389, 2014.
- [9] J. Cheng, L. Li, C. Li, Y. Liu, A. Liu, R. Qian, and X. Chen, "Remove diverse artifacts simultaneously from a single-channel EEG based on SSA and ICA: A semi-simulated study," *IEEE Access*, vol. 7, pp. 60 276–60 289, 2019.
- [10] W. Sun, Y. Su, X. Wu, and X. Wu, "A novel end-to-end 1D-ResCNN model to remove artifact from EEG signals," *Neurocomputing*, vol. 404, pp. 108–121, 2020.
- [11] H. Zhang, M. Zhao, C. Wei, D. Mantini, Z. Li, and Q. Liu, "EEGdenoiseNet: A benchmark dataset for end-to-end deep learning solutions of EEG denoising," *arXiv preprint arXiv:2009.11662*, 2020.
- [12] H. Zhang, C. Wei, M. Zhao, Q. Liu, and H. Wu, "A novel convolutional neural network model to remove muscle artifacts from EEG," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1265–1269.
- [13] J. Yu, C. Li, K. Lou, C. Wei, and Q. Liu, "Embedding decomposition for artifacts removal in EEG signals," *Journal of Neural Engineering*, vol. 19, no. 2, p. 026052, 2022.
- [14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [15] A. Mert and A. Akan, "Hilbert-Huang transform based hierarchical clustering for EEG denoising," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.
- [16] V. Mihajlović, S. Patki, and B. Grundlehner, "The impact of head movements on EEG and contact impedance: An adaptive filtering solution for motion artifact reduction," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 5064–5067.
- [17] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv*, 2017.
- [19] Q. Hou, C.-Z. Lu, M.-M. Cheng, and J. Feng, "Conv2former: A simple transformer-style convnet for visual recognition," *arXiv preprint arXiv:2211.11943*, 2022.
- [20] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, "Intriguing findings of frequency selection for image deblurring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1905–1913.
- [21] Y. Dai, F. Giesecke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 3560–3569.
- [22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.
- [23] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck, *Discrete-time signal processing*. Upper Saddle River, N.J.: Prentice Hall, 1999.
- [24] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
- [25] M. A. Klados and P. D. Bamidis, "A semi-simulated eeg/eog dataset for the comparison of eeg artifact rejection techniques," *Data in Brief*, vol. 8, pp. 1004–1006, 2016.
- [26] M. G. Terzano, L. Parrino, A. Smerieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa *et al.*, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep," *Sleep Medicine*, vol. 3, no. 2, pp. 187–199, 2002.
- [27] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "Isruc-sleep: A comprehensive public dataset for sleep researchers," *Computer Methods and Programs in Biomedicine*, vol. 124, pp. 180–192, 2016.
- [28] Head lateral anatomy. Dec. 23, 2006. [Online]. Available: <https://homes.esat.kuleuven.be/~biomed/biosource/biosource.htm>
- [29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [30] X. Chen, Q. Liu, W. Tao, L. Li, S. Lee, A. Liu, Q. Chen, J. Cheng, M. J. McKeown, and Z. J. Wang, "ReMAE: User-friendly toolbox for removing muscle artifacts from EEG," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 2105–2119, 2019.
- [31] P. Gajbhiye, N. Mingchinda, W. Chen, S. C. Mukhopadhyay, T. Wilaiprasitporn, and R. K. Tripathy, "Wavelet domain optimized Savitzky–Golay filter for the removal of motion artifacts from EEG recordings," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.
- [32] I. I. Goncharova, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "EMG contamination of EEG: spectral and topographical characteristics," *Clinical Neurophysiology*, vol. 114, no. 9, pp. 1580–1593, 2003.
- [33] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 733–13 742.
- [34] X. Ding, H. Chen, X. Zhang, J. Han, and G. Ding, "Repmlpnet: Hierarchical vision mlp with re-parameterized locality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 578–587.