EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization

Yonghao Song, *Graduate Student Member, IEEE*, Qingqing Zheng, *Member, IEEE*, Bingchuan Liu, *Student Member, IEEE*, and Xiaorong Gao, *Member, IEEE*

***Abstract*— Due to the limited perceptual field, convolutional neural networks (CNN) only extract local temporal features and may fail to capture long-term dependencies for EEG decoding. In this paper, we propose a compact Convolutional Transformer, named EEG Conformer, to encapsulate local and global features in a unified EEG classification framework. Specifically, the convolution module learns the low-level local features throughout the one-dimensional temporal and spatial convolution layers. The self-attention module is straightforwardly connected to extract the global correlation within the local temporal features. Subsequently, the simple classifier module based on fully-connected layers is followed to predict the categories for EEG signals. To enhance interpretability, we also devise a visualization strategy to project the class activation mapping onto the brain topography. Finally, we have conducted extensive experiments to evaluate our method on three public datasets in EEG-based motor imagery and emotion recognition paradigms. The experimental results show that our method achieves state-of-the-art performance and has great potential to be a new baseline for general EEG decoding. The code has been released in https://github.com/eeyhsong/EEG-Conformer.**

***Index Terms*— EEG classification, self-attention, transformer, brain-computer interface (BCI), motor imagery.**

I. Introduction

BRAIN-COMPUTER interface (BCI) is an emerging technology in recent decades, which establishes a direct pathway between external devices and the brain. BCI has brought many new applications in motor rehabilitation, emotion recognition, human-machine interaction, etc [1], [2], [3]. Among various non-invasive techniques, electroencephalograph (EEG) is widely employed to detect neural activities, using a cap with multiple electrodes to capture changes in potential on the scalp. With collected EEG signals, people can decode them into movement, vision, and other intentions, then use the results to control external devices such as computers, wheelchairs, and robots [4], [5], [6]. Although EEG is convenient and low-cost, EEG decoding is still very challenging due to many artifacts caused by impedance and other physiological signals [7].

Various pattern recognition methods have been developed to decode useful information from noisy EEG signals. These methods extract features and perform classification for different tasks. For example, common spatial pattern (CSP) is used to enhance spatial features for motor imagery (MI) tasks [8]. The filter bank is further embedded for frequency rhythms in MI and steady-state visually evoked potential (SSVEP) classification [9]. Continuous wavelet transform (CWT) is utilized to extract time-frequency features from EEG signals for detecting dementia [10]. Empirical wavelet transform (EWT) is applied to obtain improved time-frequency features from EEG with good performance for seizure detection [11], [12]. With these representative features, we can effectively achieve EEG decoding just by following a classifier, such as support vector machine (SVM) and multi-layer perceptron (MLP) [13], [14]. However, most traditional feature extraction methods are task-dependent, meaning that features are obtained with specific prior knowledge for different BCI paradigms and of limited generalization. Moreover, optimizing feature extraction and classifier separately may also lead to imperfect global optimization.

Researchers further attempt to decode EEG with end-to-end convolutional neural network (CNN), which has shown excellent representation capability in computer vision tasks [15]. As expected, the modified CNN model, ConvNet [16], achieves comparable performance to traditional algorithms on EEG classification tasks, learning discriminative features in convolutional layers. Similarly, the compact EEGNet [17] demonstrates remarkable temporal feature perception and shows good generalization across multiple BCI paradigms. Nevertheless, due to the limited kernel size, CNNs learn features with local receptive fields, but fail to acquire long-term dependencies that are crucial for time series. Recurrent neural networks (RNN) and long short-term memory (LSTM) are further proposed to capture temporal features for EEG classification [18], [19]. However, such models cannot be trained in parallel, and the dependency influence computed

Manuscript received 20 September 2022; revised 21 November 2022; accepted 11 December 2022. Date of publication 16 December 2022; date of current version 2 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U2241208, Grant 62206270, and Grant 62171473; in part by the Key Research and Development Program of Ningxia under Grant 2022CMG02026; in part by the GuangDong Basic and Applied Basic Research Foundation under Grant 2021A1515110598; and in part by the Doctoral Brain+X Seed Grant Program of Tsinghua University. *(Corresponding author: Xiaorong Gao.)*

Yonghao Song, Bingchuan Liu, and Xiaorong Gao are with the Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China (e-mail: gxr-dea@tsinghua.edu.cn).

Qingqing Zheng is with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.

Digital Object Identifier 10.1109/TNSRE.2022.3230250

by the hidden states is quickly lost after a few time steps.

Lately, attention-based Transformer models have made waves in natural language and image processing due to the inherent perception of global dependencies [20]. Transformers also emerge in EEG decoding and achieve good performance, by leveraging long-term temporal relationships [21], [22]. However, such models ignore learning local features, which are also necessary for EEG decoding. In that case, extra feature extraction processing, such as activity map and spatial filter, has to be added for compensation [23], [24]. And there is no detailed analysis and visualization to clarify how Transformer works for EEG decoding. Therefore, Transformer models remain explored in the EEG domain and not yet capable of serving as end-to-end backbones for raw EEG classification.

To tackle the above issues, we propose a Convolutional Transformer framework, named EEG Conformer, to comprehensively exploit the advantages of both CNN and Transformer. The overall framework consists of three components in series, namely, the convolution module, the self-attention module and the classifier. In the convolution module, we first employ temporal and spatial convolutions to capture local temporal and spatial features, respectively. An average pooling layer is followed to slice temporal feature segments, which not only reduces the model complexity but also removes redundant information. Then, we treat all convolutional channels at each point in the time dimension as a token and feed them into the self-attention module, which further learns the global temporal dependencies with self-attention layers. Finally, simple fully-connected layers are used to obtain the decoding results. Detailed comparative experiments are performed on several EEG datasets of different paradigms to reveal the remarkable performance of EEG Conformer.

The contributions are summarized as follows:

- We propose a concise network named Convolutional Transformer (EEG Conformer) to couple local features and global features of EEG signals. It achieves state-of-the-art results on three public datasets, with the potential to be a new backbone for EEG decoding.
- We conduct extensive experiments to investigate the effect of the Transformer module and attention parameters. The results show that our model is insensitive to the depth and head number of the self-attention module while processing EEG data.
- We design a novel visualization based on class activation mapping and topography to illustrate how the model learns essential features from a global perspective.

The rest of this paper is organized as follows. See Section II for the related works. A detailed description of the method is given in Section III. We present experiments and results in Section IV. After then, there is a careful discussion in Section V. Finally, we draw a conclusion in Section VI.

II. Related Works

A. EEG Decoding With Machine Learning

Advances in machine learning have facilitated the development of EEG classification [25], [26], [27]. In recent years, end-to-end deep learning methods have been widely adopted to process EEG signals and show good generalization. Schirrmeister et al. [16] proposed a shallow ConvNet with temporal and spatial convolutional layers to decode task-related information from raw EEG signals. Similarly, Lawhern et al. [17] developed a compact EEGNet with convolution along the temporal dimension and depthwise convolution along the spatial dimension, respectively. These two robust EEG-based CNN backbones soon inspired many excellent studies. Sakhavi et al. [28] used CNN to learn temporal information from the filter bank CSP features and select architecture parameters for each subject. Shan et al. [29] leveraged the cross-channel topological connectivity by introducing graphs to spatial-temporal CNN. Hong et al. [30] extracted subject-invariant features via CNN in an adversarial learning-driven domain adaptation framework. There are also works that proposed some tricks to enhance the performance of CNN for EEG-based motor imagery tasks [31], [32].

B. Attention-Based Transformer Network

Attention-based Transformers derived from machine translation have attracted much attention. The attention mechanism has the intrinsic ability to evaluate global dependencies on very long sequences [20]. Dosovitskiy et al. [33] applied pure Transformer on image patches and achieved good results compared with CNN-based methods. Transformers are brought into EEG processing because the global interaction is non-negligible in task-related EEG trials. Kostas et al. [34] designed a pre-training and fine-tuning approach using Transformer for EEG classification tasks. Song et al. performed feature learning from the spatial and temporal domains, where the EEG signal was sliced along the time dimension [22]. A similar framework was given by Liu et al. [35] to deal with differential entropy features of EEG. Bagchi et al. [23] converted EEG to multi-frame activity maps, then used a CNN-based module as well as combined CNN and Transformer modules to capture useful information. However, feature extraction reduces the information in raw data and often tends to depend on specific tasks. And previous studies usually focused on how to improve EEG decoding accuracy, while neglecting to interpret the role of global features with long-term dependencies visually. Therefore, inspired by the works above, we propose the EEG Conformer as an efficient backbone with novel visualization.

III. Methods

A. Overview

As an emerging neural network, Transformer is good at capturing global dependencies, but how to effectively apply it in EEG decoding remains to be explored. In this paper, we propose a novel framework, called EEG Conformer, to combine CNN and Transformer straightforwardly for end-to-end EEG classification. Borrowing ideas from CNN and Transformer, the Conformer uses convolution to learn local temporal and spatial features and then adopts self-attention to encapsulate global temporal features.

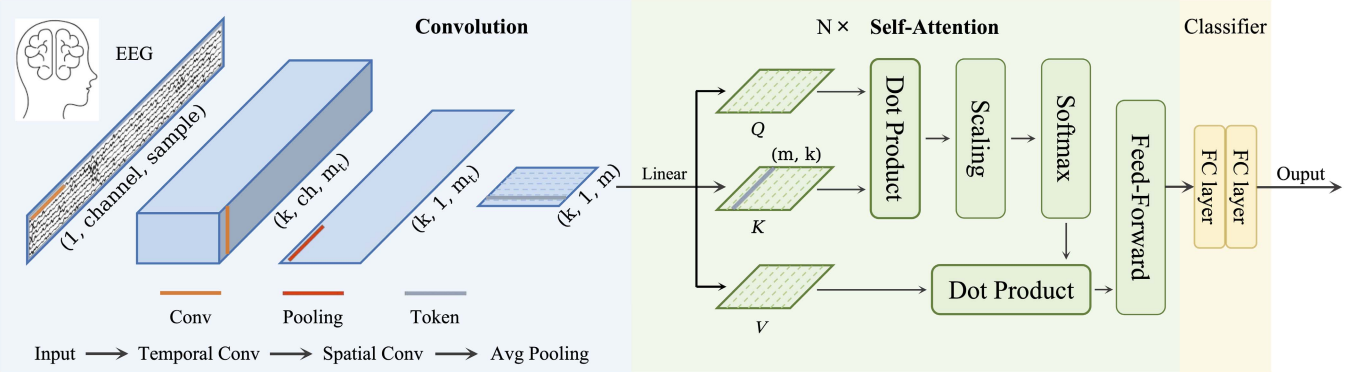

Fig. 1. The framework of Convolutional Transformer (Conformer), including a convolution module, a self-attention module, and a classifier module.

The overall framework is depicted in Fig. 1. The architecture comprises three components: a convolution module, a self-attention module, and a fully-connected classifier. In the convolution module, taking the raw two-dimensional EEG trials as the input, temporal and spatial convolutional layers are applied along the time dimension and electrode channel dimensions, respectively. Then, an average pooling layer is utilized to suppress noise interference while improving generalization. Secondly, the spatial-temporal representation obtained by the convolution module is fed into the self-attention module. The self-attention module further extracts the long-term temporal features by measuring the global correlations between different time positions in the feature maps. Finally, a compact classifier consisting of several fully-connected layers is adopted to output the decoding results.

B. Preprocessing

The raw EEG trials are of size $ch \times sp$, where ch represents electrode channels and sp denotes time samples. Without introducing additional task-dependent prior knowledge, we only use a few steps to pre-process the raw EEG data. First, band-pass filtering is employed to filter out extraneous high and low-frequency noise. Here, we use a 6-order Chebyshev filter to preserve task-relevant rhythms. Then, a Z-score standardization is performed to reduce the fluctuation and nonstationarity as

$$x_o = \frac{x_i - \mu}{\sqrt{\sigma^2}}, \tag{1}$$

where x_i and x_o denote band-pass filtered data and the output of standardization, respectively. μ and σ^2 represent the mean and variance, calculated with the training data and used directly for the test data.

C. Network Architecture

As shown in Fig. 1, EEG Conformer consists of three steps in the end-to-end process: convolution module, self-attention module, and fully-connected classifier. The input is a batch of pre-processed EEG trials with channel and sample dimensions, expanded by one dimension as the convolution channel. The output is the probability of different EEG categories.

TABLE I
NETWORK ARCHITECTURE OF THE CONVOLUTION MODULE

Layer	In	Out	kernel	stride
Temporal Conv	1	k	(1, 25)	(1, 1)
Spatial Conv	k	k	(ch, 1)	(1, 1)
Avg Pooling	k	k	(1, 75)	(1, 15)
Rearrange	(k, 1, m) → (m, k)			

1) Convolution Module: Inspired by [16] and [17], we design the convolution module by separating the two-dimensional convolution operator into two one-dimensional temporal and spatial convolution layers. The first layer has k kernels of size $(1, 25)$ with a stride of $(1, 1)$, which means the convolution is performed over the time dimension. The second layer keeps k kernels of size $(ch, 1)$ with a stride of $(1, 1)$, where ch equals the number of electrode channels of EEG data. This layer acts as a spatial filter to learn the representation of the interactions between different electrode channels. Subsequently, batch normalization is adopted to boost the training process and alleviate overfitting. We use exponential linear units (ELUs) as the activation function for nonlinearity following [17]. The third layer is an average pooling along time dimension with the kernel size of $(1, 75)$ and a stride of $(1, 15)$. This pooling layer smooths the temporal features, which not only avoids overfitting, but also reduces the computational complexity. As shown in Table I, the hyper-parameter k is set to 40. In the end, we rearrange the feature maps of the convolution module, squeeze the electrode channel dimension, and transpose the convolution channel dimension with the time dimension. In this way, we feed all feature channels of each temporal point as a token into the next module.

2) Self-Attention Module: We assume that the context-dependent representation within the low-level temporal-spatial features would benefit the EEG decoding, because the neural activities are coherent. In this module, we use self-attention to learn global temporal dependencies of EEG features, complementing the limited receptive field in the convolution module. The arranged tokens from the previous module are linearly transformed into equal-shaped triplicates, called query (Q), key (K), and value (V). Dot product is employed over Q and K to evaluate the correlation between different tokens. A scaling factor is designed to avoid vanishing gradients,

thus ensuring stable training. The result is passed through a $Softmax$ function to obtain the weighting matrix, namely the attention score. Then the attention score is weighted on V with a dot product [20]. This process can be formulated as

$$\text{Attention}(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{k}})V, \tag{2}$$

where k denotes the length of a token. Besides, two fully-connected feed-forward layers are connected behind to enhance the fitting ability. The input and output sizes of this process remain the same. The entire attention computation is repeated N times in the self-attention module.

We also employ the multi-head strategy to further improve representation diversity. The tokens are equally divided into h segments and fed into the self-attention module separately, and the results are concatenated as the module output [20]. The process can be expressed as

$$\begin{aligned} \text{MHA}(Q, K, V) &= [head_0; \cdots; head_{h-1}], \\ head_l &= \text{Attention}(Q_l, K_l, V_l) \end{aligned} \tag{3}$$

where MHA stands for multi-head attention, Q_l, K_l, $V_l \in \mathbb{R}^{m \times k/h}$ denote the query, key, and value obtained by linear transformation of divided token in the l-th head, respectively.

3) Classifier Module: Finally, we adopt two fully-connected layers as the classifier module, which outputs an M-dimensional vector after $Softmax$ function. Cross-entropy is used as the loss function of the whole framework as

$$\mathcal{L} = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{c=1}^{M} y \log(\hat{y}). \tag{4}$$

where M represents the number of EEG categories, y and $\hat{y}$ are the ground truth and predicted label, respectively. N_b denotes the number of trials in a batch.

To sum up, the band-pass filtered and standardized EEG data are fed into the model firstly. Then the data are sequentially passed through the temporal and spatial convolution layers and arranged into tokens by the pooling layer. After that, N self-attention layers are used, followed by fully-connected layers to output the classification results.

IV. Experiments and Results

In this section, we conduct experiments to verify the proposed network on three public EEG datasets, including popular motor imagery and emotion recognition paradigms. We not only compare our method with different state-of-the-art approaches, but also demonstrate the improvements by introducing the attention-based Transformer through ablation studies. We also present detailed comparative experiments to show the influence of attention parameters on overall performance. Finally, we design different visualization methods for interpretability.

A. Datasets

We evaluate our method on three widely used EEG datasets, including BCI competition IV dataset 2a,[1] BCI competition IV dataset 2b,[2] SEED[3] [36] These EEG datasets were collected with different acquisition devices, paradigms, number of subjects, and sample size, thus fairly validating the generalization of our method.

1) Dataset I: BCI Competition IV Dataset 2a provided by Graz University of Technology consists of EEG data from 9 subjects. There were four motor imagery tasks, covering the imagination of moving left hand, right hand, both feet, and tongue. Two sessions on different days were collected with twenty-two Ag/AgCl electrodes at a sampling rate of 250 Hz. One session contained 288 EEG trials, i.e., 72 trials per task. We used $[2, 6]$ seconds of each trial and filtered the EEG data to $[4, 40]$ Hz with a band-passed filter as [8] in our experiments. The first session was used for training and the second session for test.

2) Dataset II: BCI Competition IV Dataset 2b provided by Graz University of Technology consists of EEG data from 9 subjects. There were two motor imagery tasks, covering the imagination of moving left and right hand. Five sessions were collected with three bipolar electrodes (C3, Cz, and C4) at a sampling rate of 250 Hz and each session contained 120 trials. We used the $[3, 7]$ seconds of each trial in the experiments. We also performed band-pass filtering between $[4, 40]$ Hz to reduce high and low-frequency noise. The first three sessions were training set, and the last two sessions were test set.

3) Dataset III: SEED dataset provided by Shanghai Jiao Tong University consists of emotion-based EEG signals from 15 subjects. There were three emotions, including positive, neutral, and negative, stimulated by fifteen film clips. The data collection process was repeated three times on each subject at approximately weekly intervals. The EEG signals were captured with 62 electrodes at a sample rate of 1000 Hz and subsequently downsampled to 200 Hz. Each sample was segmented with a non-overlapped one-second time window, resulting in a total of 3394 trials from one session. We also performed band-pass filtering of $[4, 47]$ Hz on the data. Five-fold cross-validation was used in the SEED dataset.

B. Data Augmentation

EEG acquisition is time-consuming, which results in small datasets that are prone to overfitting. Some methods employ data augmentation to feed enough samples into the models [16]. However, the conventional strategies of adding Gaussian noise or cropping may further lower the signal-to-noise ratio or destroy the original coherence. Therefore, we employ segmentation and reconstruction (S&R) in the time domain to generate new data. Follow [37], the training samples of the same category are equally divided into N_s segments, then randomly concatenated while maintaining the original time order. We generate the augmented data of the same size as the batch in each iteration.

C. Experiment Details

Our method is implemented with PyTorch library in Python 3.10 with a Geforce 3090 GPU. We train the model using

[1] https://www.bbci.de/competition/iv/desc_2a.pdf

[2] https://www.bbci.de/competition/iv/desc_2b.pdf

[3] https://bcmi.sjtu.edu.cn/home/seed/seed.html

Adam optimizer with the learning rate, β_1 and β_2 of 0.0002, 0.5, and 0.999, respectively. We set the execution times N of self-attention to 6, the number of heads h to 10, and the N_s in S&R to 8. The classification accuracy and kappa are used as evaluation metrics for the overall performance. Kappa can be calculated with

$$kappa = \frac{p_o - p_e}{1 - p_e}, \tag{5}$$

where p_o represents the average accuracy of all the trials and p_e denotes the accuracy of random guesses. Wilcoxon Signed-Rank Test is employed to analyze the statistical significance.

D. Baseline Comparison

We conduct extensive subject-dependent experiments and compare our method with some state-of-the-art approaches on three public datasets.

Datasets I is currently the most widely used multi-class motor imagery dataset. We compare many representative methods, which have achieved impressive performance on this dataset. For example, FBCSP [8], the winner of BCI Competition IV using hand-crafted spatial features; ConvNet [16] and EEGNet [17], which have shown remarkable results on many EEG datasets with CNN-based end-to-end frameworks; C2CM [28], which inputs the FBCSP features to the CNN model, combining the advantages of traditional feature extraction and deep learning methods; FBCNet [38], extracting spectro-spatial features by spatial filtering multi-view data. We even compare with deep representation-based domain adaptation (DRDA) [39] that utilizes data from other subjects for enhancement with adversarial learning.

The classification performance of each subject and the average results on Dataset I are presented in Table II. We can observe that our Conformer significantly improves the accuracy by 10.91% over FBCSP ($p < 0.01$), which depends on traditional feature extraction. The results also show that other deep learning methods, such as ConvNet and EEGNet, outperform FBCSP, indicating that the CNN-based methods have strong feature representation capability. However, these CNN-based methods only focus on local features due to the limited perceptual field, and ignore the global correlation, which may compromise the decoding accuracy for coherent EEG series. Differently, our method encapsulates both the local and global dependencies by integrating Transformer architecture on the basis of the original CNN. Thus, Conformer obtains better results on most subjects and achieves significant upgrades on average accuracy ($p < 0.05$) and kappa. C2CM and FBCNet effectively combine the idea of hand-crafted features and deep models, but still cannot beat ours except for subject 5 ($p < 0.05$), although C2CM fine-tuned the model parameters for each subject. DRDA brings in data from other subjects with the distribution aligned to the target subject, which is still inferior to ours just using the data of target subject ($p < 0.05$), once again demonstrating the effectiveness of leveraging both local and global features.

Then we present the comparison with several state-of-the-art methods on Dataset II in Table III. We can see that the binary classification results show similar trends as in Dataset I.

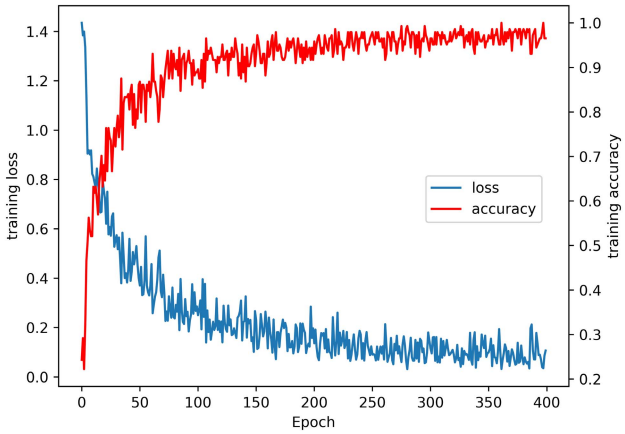

Fig. 2. Loss and accuracy during training of EEG Conformer.

Conformer promotes the overall performance significantly compared with FBCSP ($p < 0.05$), with even an increasing accuracy of 12.5% on subject 1. There is an obvious boost by contrast with other end-to-end methods using just CNN architecture, with improvements of 5.25% and 4.15% for ConvNet ($p < 0.05$) and EEGNet ($p < 0.01$). The average accuracy and kappa of our method still precede DRDA on almost all the subjects, which further validates the efficacy of our method.

We also comprehensively evaluate our method on Dataset III of multi-category EEG emotion data. We compare with machine learning methods like SVM [36], which first achieved notable results on this dataset; graph regularized extreme learning machine (GELM) [40] with a single feed-forward layer to learn discriminative features, and regions to global spatial-temporal neural network (R2G-STNN) [42] that adopts the bidirectional long short term memory to learn spatial and temporal features of emotion EEG signals. Besides, graph-based neural networks learning the intrinsic relationship among different EEG channels such as dynamical graph convolutional neural network (DGCNN) [41] and regularized graph neural network (RGNN) [43] are also included for comparison. The results are presented in Table IV. It can be seen that Conformer is still competitive on Dataset III compared with other state-of-the-art methods. In this way, our method achieves impressive performance on both motor imagery and emotion recognition paradigms, illustrating that our method has good generalization.

E. Training Process

In image processing, Transformer models often need a large amount of data for pre-training to achieve good results in downstream tasks. However, pre-training is not used in EEG Conformer, due to the limited data for calibration. We demonstrate the trend of loss and accuracy during training in Fig. 2. The process is stable under the lightweight use of the self-attention module. It can be noticed that the model converges quickly around the 250_{th} epoch. Moreover, our method is also efficient. We train the Conformer model continuously with the first subject in Dataset I for 2000 epochs

TABLE II
COMPARISONS WITH STATE-OF-THE-ART METHODS ON DATASETS I

datasets	methods	S01	S02	S03	S04	S05	S06	S07	S08	S09	average	kappa
I	FBCSP [8]	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75	0.5700
	ConvNet [16]	76.39	55.21	89.24	74.65	56.94	54.17	92.71	77.08	76.39	72.53	0.6337
	EEGNet [17]	85.76	61.46	88.54	67.01	55.90	52.08	89.58	83.33	86.81	74.50	0.6600
	C2CM [28]	87.50	**65.28**	90.28	66.67	62.50	45.49	89.58	83.33	79.51	74.46	0.6595
	FBCNet [38]	85.42	60.42	90.63	76.39	**74.31**	53.82	84.38	79.51	80.90	76.20	0.6827
	DRDA [39]	83.19	55.14	87.43	75.28	62.29	57.15	86.18	83.61	82.00	74.74	0.6632
	Conformer	**88.19**	61.46	**93.40**	**78.13**	52.08	**65.28**	**92.36**	**88.19**	**88.89**	**78.66**	**0.7155**

TABLE III
COMPARISONS WITH STATE-OF-THE-ART METHODS ON DATASETS II

datasets	methods	S01	S02	S03	S04	S05	S06	S07	S08	S09	average	kappa
II	FBCSP [8]	70.00	60.36	60.94	97.50	93.12	80.63	78.13	92.50	86.88	80.00	0.6000
	ConvNet [16]	76.56	50.00	51.56	96.88	**93.13**	85.31	83.75	91.56	85.62	79.37	0.5874
	EEGNet [17]	75.94	57.64	58.43	98.13	81.25	88.75	84.06	93.44	89.69	80.48	0.6096
	DRDA [39]	81.37	62.86	63.63	95.94	93.56	88.19	85.00	**95.25**	90.00	83.98	0.6796
	Conformer	**82.50**	**65.71**	**63.75**	**98.44**	86.56	**90.31**	**87.81**	94.38	**92.19**	**84.63**	**0.6926**

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS ON DATASETS III

datasets	methods	accuracy	kappa
III	SVM [36]	86.08	0.7912
	GELM [40]	91.07	0.8661
	DGCNN [41]	90.40	0.8560
	R2G-STNN [42]	93.38	0.9007
	RGNN [43]	94.24	0.9136
	Conformer	**95.30**	**0.9295**

on a single GPU, obtaining an average time of 0.27 seconds per epoch.

F. Ablation Study

The key improvement of EEG Conformer over the CNN-based approach is the addition of the attention-based Transformer module for learning global representations. As well, data augmentation may have contributed to the final results. Therefore, We conduct an ablation study on Dataset I, as shown in Fig. 3, where the self-attention module and the S&R data augmentation is removed separately. It can be seen that when the Transformer part is removed, there is a substantial decrease in the result on each subject. Subject 6 reduces the most by 8.68%, and subject 3 reduces the least by 3.12%. The average accuracy drops significantly by 6.02% ($p < 0.01$). Similar to ConvNet [16], the experimental results in Fig. 3 also show the data augmentation strategy can help improve the performance of our model. The overall performance improves by an average accuracy of 3.75% ($p < 0.01$) compared with the one without data augmentation. Interestingly, the improvement is only 1.04% for subject 1 with better discrimination, while for subject 5 and 6, who perform originally poor, the improvements are more significant and reach 4.86% and 5.56%, respectively. Therefore, the introduction of data augmentation in the training process enhances the robustness of Conformer.

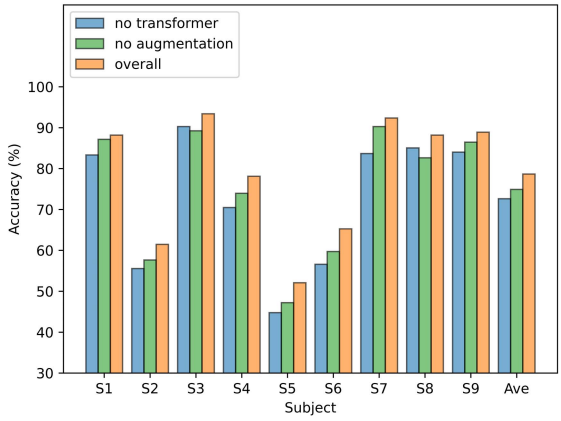

Fig. 3. Ablation study on the self-attention module and data augmentation.

G. Parameter Sensitivity

In this section, we evaluate in detail the impact of several important parameters in the self-attention module on performance. These include the depth N of self-attention layers, the number h of attention heads, and the design of the pooling kernel, which constructs the input for learning global features.

Depth is usually a crucial factor affecting the fitting ability of end-to-end models, such as CNN and Transformer. As in Fig. 4, we explore the effect of depth on EEG Conformer by gradually increasing the layers of self-attention module from 0 to 15. It can be seen that for Dataset I, there is a significant improvement in accuracy when the depth goes from 0 to 1 ($p < 0.01$). It illustrates the introduction of Transformer does help EEG decoding once again. For the other depths, the highest accuracy is only 1.24% higher than the lowest. And the difference is not significant ($p > 0.05$). However, as shown in the parameter curves, the number

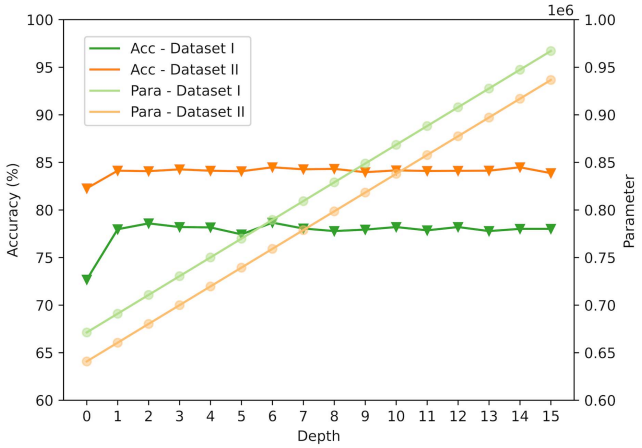

Fig. 4. The influence of the depth of the self-attention module (from 0 to 15) on the accuracy and the amounts of parameters for Dataset I and II.

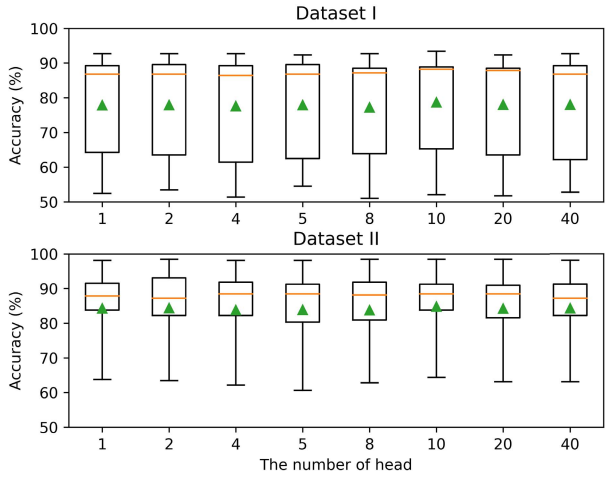

Fig. 5. The influence of the number of attention heads on the accuracy for different datasets.

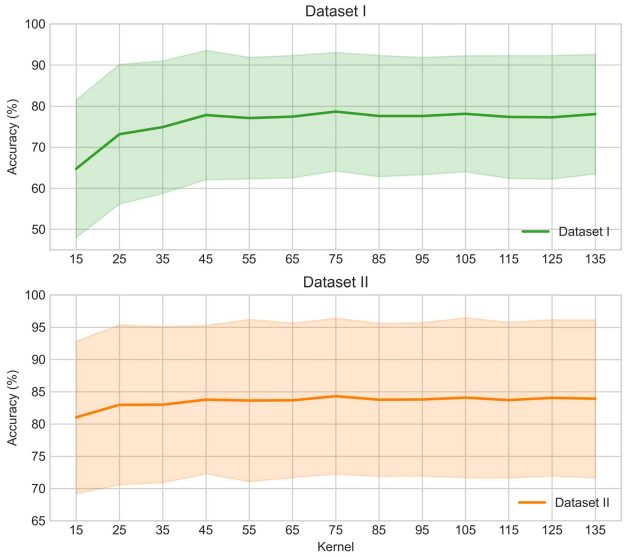

Fig. 6. The influence of different kernel sizes in the pooling layer, namely, the token size of the self-attention module.

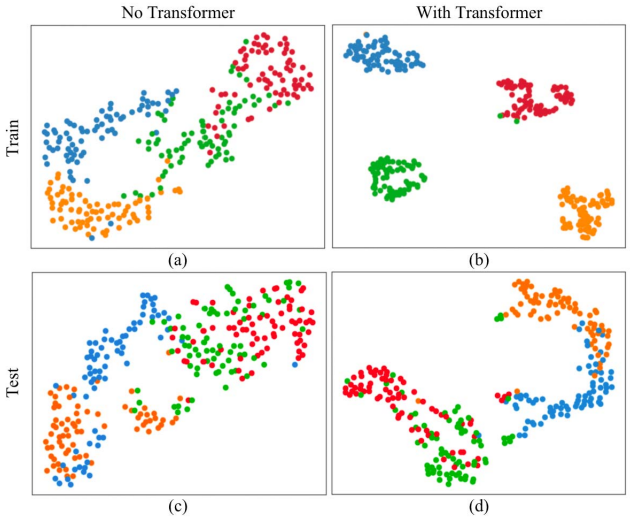

Fig. 7. t-SNE visualization illustrates the significance of introducing Transformer for feature learning. Different colors represent different categories.

of parameters increases proportionally with depth, which makes the model less cost-effective. The same evaluation on Dataset II also shows the insensitivity of Conformer to self-attention depth.

Head is an important parameter of common Transformer models based on the multi-head attention. It is reported that it can help to learn different aspects of features. We also compare the impact of different head selections on the model, as shown in Fig. 5, choosing eight head numbers between 1 and 40. From the box, there is no clear pattern for the effect of different head numbers on the results. The distribution of different subjects has no obvious difference. The average accuracy maintains a mild fluctuation, where the range is just 1.43% on Dataset I and 1.02% on Dataset II. The performance has a slight upward trend as the head number increases but then declines. The average accuracy is 0.82% higher in Dataset I and 0.50% higher in Dataset II ($p > 0.05$), when the number of heads is taken as 10 than when it is taken as 1. Overall, changes in the number of heads have not yet shown a significant effect in prompting feature learning.

The token determined by the pooling kernel, is also a critical factor for the self-attention module. If the kernel size is too large, the temporal features would be too smoothed and lose useful details. Thus, it is difficult for the model to perceive the global relationship between details. In contrast, if the kernel is too small, the performance may be easily affected by local noise. We compare the effect of different pooling kernel choices on model performance as in Fig. 6. The kernel size is taken from 15 to 135 with an interval of 10. It is clear to see a substantial upgrade in the average accuracy when the kernel size starts to grow. A gain of 13.08% ($p < 0.01$) is obtained on Dataset I by increasing the kernel from 15 to 45. After that, the results flatten out and do not rise observably with increasing kernel size. The experiments demonstrate that applying self-attention to sufficiently large slices does make sense for EEG with a low signal-to-noise ratio.

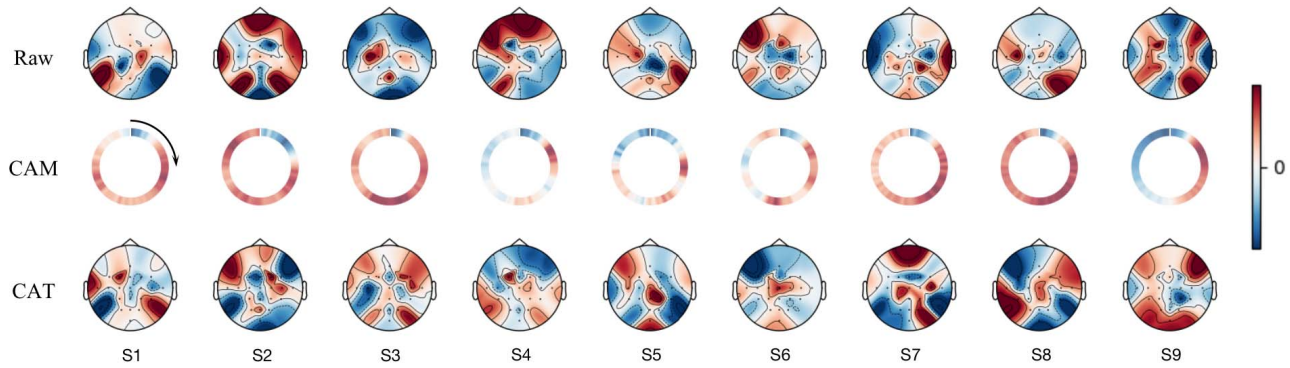

Fig. 8. Raw EEG topography averaged over all trials of each subject, Class Activation Mapping (CAM) of the Transformer module on the input EEG, Class Activation Topography (CAT) we designed to show CAM-weighted EEG. Raw shows that many regions are activated throughout the trial. CAM shows that our model pays different attention to different ranges in the time domain. CAT shows our model focus on areas of the motor cortex in motor imagery data.

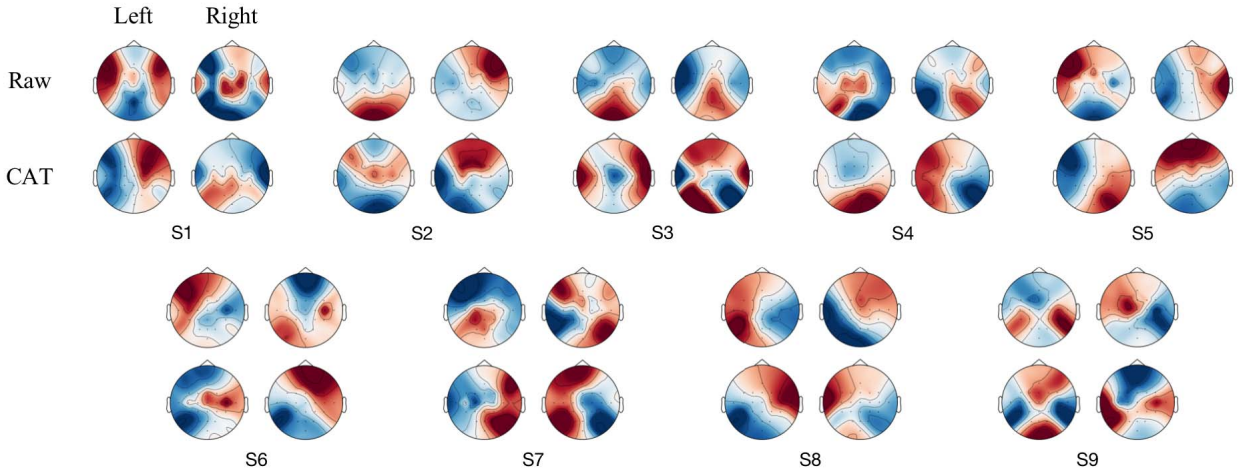

Fig. 9. CAT shows the ERD/ERS phenomena on both the data of imagining left and right hand movements, compared to the irregular patterns in raw EEG topography. Contralateral activation and ipsilateral inhibition can be clearly observed in the CAT of several subjects, such as S1, S7, and S8.

H. Visualization

We visualize two perspectives to show the interpretability of EEG Conformer, including deep features by t-SNE [44] and spatial-temporal features reflected on topography.

1) Feature Distribution: t-distributed stochastic neighbor embedding (t-SNE) is a popular statistical dimension reduction and visualization method. The feature distribution of Subject 1 in Dataset I after adequate training with and without Transformer is shown in Fig. 7. We can see that for training data, the features of different categories are relatively close without the help of Transformer. After adding Transformer, the inter-category distance becomes larger, and the intra-category distance becomes smaller, as in Fig. 7(b). On the other hand, the aliasing between categories is evident in the absence of Transformer, which sharpens category boundaries in Fig. 7(d).

2) Global Representation: Transformer is introduced to learn global temporal dependencies in EEG data, which means locating more important information for decoding tasks from time series. We use topography and Gradient-weighted Class Activation Mapping (CAM) [45] to show the global representation learned by our model with motor imagery Dataset I in Fig. 8. The first row in the figure denotes that all training trials of each subject are averaged for the topography. There are no apparent clues of active brain regions among different subjects. CAM is adopted to monitor the time period that the self-attention module pays attention to on the EEG features, as shown in the second row of Fig. 8. EEG data is drawn as a circle, clockwise from the top during the motor imagery process. Different activation is presented at different time. As expected, data of all subjects are attenuated at the beginning of trials, which may indicate a latency for movement intention.

We further propose a new visualization method applied to EEG named Class Activation Topography (CAT). EEG Topography is drawn on the normalized data multiplied by the normalized CAM. From the third row of Fig. 8, most of the EEG data weighted by CAM focus on the area of the motor cortex, consistent with the paradigm of motor imagery [46]. Furthermore, the raw EEG and CAT of imagining left-hand movement and right-hand movement are plotted in Fig. 9. We are surprised to find event-related desynchronization (ERD) and event-related synchronization (ERS) phenomenon. Obvious contralateral activation and ipsilateral inhibition are observed in the CAT of several subjects, such as the first and eighth one, compared with the irregular raw EEG topography.

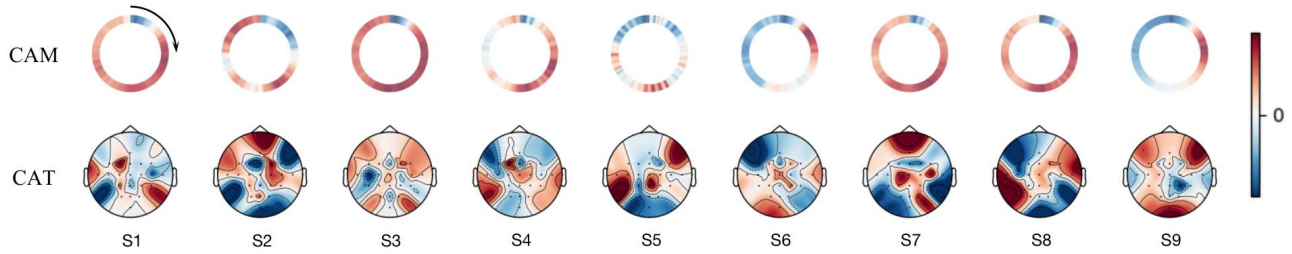

Fig. 10. CAM and CAT of the model with only 1 head in the self-attention module. The activation is close to that in Fig. 8 with 10 heads.

V. Discussion

The practicality of BCI systems depends on the performance of the decoding method. We propose a very concise but effective method named Conformer to combine the advantages of CNN and Transformer networks. Conformer is a lightweight solution for EEG decoding without pre-training. It only employs a few steps for preprocessing, including band-pass filtering and standardization, without depending heavily on specific tasks. The convolution module with both temporal and spatial convolution layers pays attention to the low-level representation, considering the local temporal features, while the self-attention module further focuses on the long-term dependencies and captures the global temporal correlation. Thus, the proposed method is capable of learning more discriminative representation compared with the existing CNN-based models.

In experiments, we can see that EEG Conformer achieves state-of-the-art results on three datasets with different paradigms and data acquisitions. The ablation study presents that Transformer module contributes significantly to the overall model, and data augmentation helps improve training performance. We also explore the effect of several key parameters on the model. The results show that the model is not sensitive to the depth and head number of the self-attention module. However, the kernel size of the pooling layer reveals a noticeable effect, which suggests that a large unit to apply attention can help to avoid the interference of local noise. Detailed visualizations are used for interpretability illustrations. The Transformer module provides better discrimination capability as the feature distribution shown by t-SNE. We also design a new visualization approach name CAT to discover the function of a layer in a model by combining EEG topography and CAM. The results demonstrate that our model focuses on changes near the motor cortex with motor imagery data. Besides, ERD/ERS produced by the imagery of left and right hands is also clearly perceived.

The role of multi-heads in the self-attention module remains unclear, so we train the model with only 1 head for Dataset I, and plot CAM and CAT in Fig. 10. We can see that the activation of the self-attention module is close to that in Fig. 8 with 10 heads. The comparison indicates that both cases learn similar global features, resulting in similar decoding accuracy. The slight difference in activation still needs to be addressed.

There are several more limitations. Firstly, we mainly validate oscillatory EEG data such as motor imagery and emotion, which lack stationary patterns as event-related potential (ERP) EEG data. Secondly, the parameter scale of the current model is not small. For Dataset I, the parameters of the Conformer increase by 17.6% compared to removing the self-attention module. These additional costs arise from the linear transformation and feed-forward layer used to calculate global dependencies. Although we have confirmed that the time cost to train the model is acceptable for actual use, it is still an issue that can be improved. Besides, the fully-connected classifier contributes a large number of parameters. Global average pooling may be used as an alternative with little performance degradation. Third, the proposed method is trained and validated on each individual, and cannot utilize useful information from other subjects. We will apply this model in ERP and subject-independent tasks in the future.

VI. Conclusion

This paper proposes a concise and efficient EEG decoding method called Conformer. Transformer is incorporated into CNN to learn global dependencies in the temporal domain. Remarkable results are achieved on different EEG datasets with detailed comparative experiments. The visualization also shows that our model locates key information that conforms to the principles of the paradigm on a global level. Overall, our model yields good performance in promoting EEG decoding.

References

[1] J. Jin et al., "A novel classification framework using the graph representations of electroencephalogram for motor imagery based brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 20–29, 2022.

[2] B. Liu, X. Chen, N. Shi, Y. Wang, S. Gao, and X. Gao, "Improving the performance of individually calibrated SSVEP-BCI by task-discriminant component analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1998–2007, 2021.

[3] J. W. Li et al., "Single-channel selection for EEG-based emotion recognition using brain rhythm sequencing," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 6, pp. 2493–2503, Jun. 2022.

[4] S. He et al., "EEG- and EOG-based asynchronous hybrid BCI: A system integrating a speller, a web browser, an E-mail client, and a file explorer," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 519–530, Feb. 2020.

[5] K.-T. Kim, H.-I. Suk, and S.-W. Lee, "Commanding a brain-controlled wheelchair using steady-state somatosensory evoked potentials," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 654–665, Mar. 2018.

[6] Y. Song, S. Cai, L. Yang, G. Li, W. Wu, and L. Xie, "A practical EEG-based human-machine interface to online control an upper-limb assist robot," *Frontiers Neurorobot.*, vol. 14, p. 32, Jul. 2020.

[7] B.-Y. Tsai, S. V. S. Diddi, L.-W. Ko, S.-J. Wang, C.-Y. Chang, and T.-P. Jung, "Development of an adaptive artifact subspace reconstruction based on Hebbian/anti-Hebbian learning networks for enhancing BCI performance," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 17, 2022, doi: 10.1109/TNNLS.2022.3174528.

[8] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, no. 1, p. 39, 2012.

[9] X. Chen, Y. Wang, S. Gao, T.-P. Jung, and X. Gao, "Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface," *J. Neural Eng.*, vol. 12, no. 4, Aug. 2015, Art. no. 046008.

[10] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia," *Neural Netw.*, vol. 123, pp. 176–190, Mar. 2020.

[11] A. Bhattacharyya, L. Singh, and R. B. Pachori, "Fourier–Bessel series expansion based empirical wavelet transform for analysis of non-stationary signals," *Digit. Signal Process.*, vol. 78, pp. 185–196, Jul. 2018.

[12] A. Bhattacharyya and R. B. Pachori, "A multivariate approach for patient-specific EEG seizure detection using empirical wavelet transform," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2003–2015, Sep. 2017.

[13] X. Wu, W.-L. Zheng, Z. Li, and B.-L. Lu, "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition," *J. Neural Eng.*, vol. 19, no. 1, Feb. 2022, Art. no. 016012.

[14] H. Göksu, "BCI oriented EEG analysis using log energy entropy of wavelet packets," *Biomed. Signal Process. Control*, vol. 44, pp. 101–109, Jul. 2018.

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

[16] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization: Convolutional neural networks in EEG analysis," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[17] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[18] S. Tortora, S. Ghidoni, C. Chisari, S. Micera, and F. Artoni, "Deep learning-based BCI for gait decoding from EEG with LSTM recurrent neural network," *J. Neural Eng.*, vol. 17, no. 4, Jul. 2020, Art. no. 046011.

[19] A. Shoeibi et al., "Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models," *Frontiers Neuroinform.*, vol. 15, Nov. 2021, Art. no. 777977.

[20] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11.

[21] J. Xie et al., "A transformer-based approach combining deep learning network and spatial–temporal information for raw EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2126–2136, 2022.

[22] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial–temporal feature learning for EEG decoding," Jun. 2021. [Online]. Available: https://arxiv.org/abs/2106.11170

[23] S. Bagchi and D. R. Bathula, "EEG-ConvTransformer for single-trial EEG-based visual stimulus classification," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108757.

[24] Y. Zheng, X. Zhao, and L. Yao, "Copula-based transformer in EEG to assess visual discomfort induced by stereoscopic 3D," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103803.

[25] F. Lotte et al., "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, Jun. 2018, Art. no. 031005.

[26] P. V. and A. Bhattacharyya, "Human emotion recognition based on time–frequency analysis of multivariate EEG signal," *Knowl.-Based Syst.*, vol. 238, Feb. 2022, Art. no. 107867.

[27] A. Bhattacharyya, R. K. Tripathy, L. Garg, and R. B. Pachori, "A novel multivariate-multiscale approach for computing EEG spectral and temporal complexity for human emotion recognition," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3579–3591, Feb. 2021.

[28] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain–computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, Nov. 2018.

[29] X. Shan, J. Cao, S. Huo, L. Chen, P. G. Sarrigiannis, and Y. Zhao, "Spatial–temporal graph convolutional network for Alzheimer classification based on brain functional connectivity imaging of electroencephalogram," *Hum. Brain Mapping*, vol. 43, no. 17, pp. 5194–5209, Jun. 2022.

[30] X. Hong et al., "Dynamic joint domain adaptation network for motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 556–565, 2021.

[31] W. Huang, W. Chang, G. Yan, Z. Yang, H. Luo, and H. Pei, "EEG-based motor imagery classification using convolutional neural networks with local reparameterization trick," *Expert Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 115968.

[32] A. M. Roy, "An efficient multi-scale CNN model with intrinsic feature integration for motor imagery EEG subject classification in brain-machine interfaces," *Biomed. Signal Process. Control*, vol. 74, Apr. 2022, Art. no. 103496.

[33] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Mar. 2022, pp. 1–21.

[34] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers Hum. Neurosci.*, vol. 15, p. 253, Jun. 2021.

[35] J. Liu, L. Zhang, H. Wu, and H. Zhao, "Transformers for EEG emotion recognition," Oct. 2021. [Online]. Available: https://arxiv.org/abs/2110.06553

[36] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[37] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proc. IEEE*, vol. 103, no. 6, pp. 871–890, Jun. 2015.

[38] R. Mane et al., "FBCNet: A multi-view convolutional neural network for brain–computer interface," Mar. 2021. [Online]. Available: https://arxiv.org/abs/2104.01233

[39] H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng, "Deep representation-based domain adaptation for nonstationary EEG classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 535–545, Feb. 2021.

[40] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul. 2019.

[41] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul. 2020.

[42] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial–temporal neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 568–578, Apr. 2022.

[43] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul. 2022.

[44] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[46] A. Schnitzler, S. Salenius, R. Salmelin, V. Jousmäki, and R. Hari, "Involvement of primary motor cortex in motor imagery: A neuromagnetic study," *NeuroImage*, vol. 6, no. 3, pp. 201–208, Oct. 1997.