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

EEG Conformer：用于脑电图解码和可视化的卷积变换器

宋永浩（IEEE 研究生会员）、郑庆庆（IEEE 会员）、刘炳川（IEEE 学生会员）和高晓蓉（IEEE 会员）

摘要：由于感知场有限，卷积神经网络 (CNN) 仅提取局部时间特征，可能无法捕捉脑电图解码的长期依赖关系。本文提出了一种紧凑的卷积变换器，称为脑电图一致性器 (EEG Conformer)，将局部和全局特征封装在统一的脑电图分类框架中。具体而言，卷积模块在一维时间和空间卷积层中学习低级局部特征。自注意力模块直接连接以提取局部时间特征中的全局相关性。随后，使用基于全连接层的简单分类器模块来预测脑电图信号的类别。为了增强可解释性，我们还设计了一种可视化策略，将类别激活映射投影到大脑拓扑图上。最后，我们在基于脑电图的运动想象和情绪识别范式中的三个公共数据集上进行了广泛的实验来评估我们的方法。实验结果表明，我们的方法达到了最佳性能，并具有成为通用脑电图解码新基准的巨大潜力。代码已发布于 https://github.com/eeyhsong/EEG-Conformer。

索引词——EEG 分类、自我注意、变压器、脑机接口 (BCI)、运动意象。

我

脑机接口 (BCI) 是近几十年来新兴的技术，它在外部设备与大脑之间建立了直接的通路。BCI 在运动康复、情绪识别、人机交互等领域带来了许多新的应用 [1], [2], [3]。在各种非侵入性技术中，脑电图 (EEG) 被广泛应用于检测神经活动，使用带有多个电极的帽子来捕捉头皮电位的变化。通过收集脑电信号，人们可以将其解码为运动、视觉和其他意图，然后利用结果来控制计算机、轮椅和机器人等外部设备[4]，[5]，[6]。尽管脑电图（EEG）方便且成本低廉，但由于阻抗和其他生理信号引起的许多伪影，脑电图解码仍然非常具有挑战性[7]。

人们已经开发出各种模式识别方法来从嘈杂的脑电信号中解码有用的信息。这些方法提取特征并对不同的任务进行分类。例如，公共空间模式 (CSP) 用于增强运动想象 (MI) 任务的空间特征 [8]。滤波器组进一步嵌入，用于 MI 中的频率节律和稳态视觉诱发电位 (SSVEP) 分类 [9]。连续小波变换 (CWT) 用于从脑电信号中提取时频特征以检测痴呆症 [10]。经验小波变换 (EWT) 用于从脑电信号中获得改进的时频特征，并具有良好的癫痫发作检测性能 [11]，[12]。利用这些代表性特征，我们只需遵循分类器，例如支持向量机 (SVM) 和多层感知器 (MLP) [13]，[14]，即可有效地实现脑电解码。然而，大多数传统的特征提取方法都是依赖于任务的，这意味着特征的获取依赖于不同 BCI 范式的特定先验知识，并且泛化能力有限。此外，分别优化特征提取和分类器也可能导致不完善的全局优化。

研究人员进一步尝试使用端到端卷积神经网络 (CNN) 解码脑电图，该网络在计算机视觉任务中表现出色 [15]，其表征能力也十分出色。正如预期，改进的 CNN 模型 ConvNet [16] 在脑电图分类任务中取得了与传统算法相当的性能，能够在卷积层中学习判别性特征。同样，紧凑型 EEGNet [17] 也展现出卓越的时间特征感知能力，并在多种 BCI 范式中表现出良好的泛化能力。然而，由于核大小有限，CNN 只能学习具有局部感受野的特征，而无法获得对时间序列至关重要的长期依赖关系。循环神经网络 (RNN) 和长短期记忆网络 (LSTM) 被进一步提出用于捕捉脑电图分类的时间特征 [18]，[19]。然而，此类模型无法并行训练，依赖关系的影响需要计算

稿件收到日期：2022 年 9 月 20 日；修改日期：2022 年 11 月 21 日；接受日期：2022 年 12 月 11 日。出版日期：2022 年 12 月 16 日；当前版本日期：2023 年 2 月 2 日。本研究部分由国家自然科学基金（批准号：U2241208、62206270 和 62171473）、宁夏重点研发计划（批准号：2022CMG02026）、广东省基础与应用基础研究基金（批准号：2021A1515110598）以及清华大学博士生“脑+X 种子”计划资助。（通讯作者：高晓蓉）

宋永浩、刘炳川和高晓蓉均就读于清华大学医学院生物医学工程系，北京 100084（电子邮件：gxr-dea@tsinghua.edu.cn）。

郑庆庆，中国科学院深圳先进技术研究院，广东省计算机视觉与虚拟现实技术重点实验室，深圳 518055。

数字对象标识符 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.

隐藏状态所获得的信息在几个时间步骤之后就会很快丢失。

最近，基于注意力机制的 Transformer 模型凭借其对全局依赖关系的固有感知，在自然语言和图像处理领域掀起了波澜 [20]。Transformer 模型也应用于脑电图解码，并通过利用长期时间关系取得了良好的性能 [21]，[22]。然而，这类模型忽略了学习局部特征，而这些特征对于脑电图解码同样必不可少。在这种情况下，需要添加额外的特征提取处理（例如活动图和空间滤波器）进行补偿 [23]，[24]。而且，目前还没有详细的分析和可视化来阐明 Transformer 模型在脑电图解码中的工作原理。因此，Transformer 模型在脑电图领域仍处于探索阶段，尚无法作为原始脑电图分类的端到端主干网络。

为了解决上述问题，我们提出了一个卷积 Transformer 框架，即 EEG Conformer，以综合利用 CNN 和 Transformer 的优势。整体框架由三个串联的组件组成，即卷积模块、自注意力模块和分类器。在卷积模块中，我们首先使用时间和空间卷积分别捕获局部时间和空间特征。随后使用平均池化层对时间特征片段进行切片，这不仅降低了模型复杂度，还消除了冗余信息。然后，我们将时间维度上每个点的所有卷积通道视为一个 token，并将其输入到自注意力模块中，该模块通过自注意力层进一步学习全局时间依赖性。最后，使用简单的全连接层获得解码结果。在多个不同范式的 EEG 数据集上进行了详细的对比实验，以揭示 EEG Conformer 的卷积性能。

贡献总结如下：

- 我们提出了一个名为卷积变换器（EEG Conformer）的简洁网络，用于耦合脑电信号的局部特征和全局特征。该网络在三个公共数据集上取得了最佳结果，有望成为脑电解码的新骨干网络。
- 我们进行了大量实验来探究 Transformer 模块和注意力机制参数的影响。结果表明，我们的模型在处理脑电数据时对自注意力模块的深度和头部数量不敏感。
- 我们设计了一种基于类激活映射和地形的新颖可视化，以说明模型如何从全局视角学习基本特征。

本文其余部分安排如下。相关工作请参见第二部分。第三部分详细描述了该方法。第四部分展示了实验和结果。之后，第五部分进行了深入的讨论。最后，第六部分得出结论。

II. RW A. 使用机器学习进行脑电图解码

机器学习的进步促进了脑电图分类的发展[25]，[26]，[27]。最近多年来，端到端深度学习方法已被广泛用于处理 EEG 信号并表现出良好的泛化能力。Schirrmeister 等人 [16] 提出了一种具有时间和空间卷积层的浅层 ConvNet，以从原始 EEG 信号中解码与任务相关的信息。同样地，Lawhern 等人 [17] 开发了一个紧凑的 EEGNet，分别沿时间维度进行卷积和沿空间维度进行深度卷积。这两个基于 EEG 的强大 CNN 主干很快激发了许多优秀的研究。Sakhavi 等人 [28] 使用 CNN 从滤波器组 CSP 特征中学习时间信息并为每个受试者选择架构参数。Shan 等人 [29] 通过将图引入时空 CNN 来利用跨通道拓扑连接。Hong 等人 [30] 在对抗学习驱动的领域自适应框架中通过 CNN 提取主题不变特征。还有一些研究提出了一些技巧来提高 CNN 在基于 EEG 的运动想象任务中的表现 [31]，[32]。

B.基于注意力机制的 Transformer 网络

源自机器翻译的基于注意力机制的 Transformer 模型备受关注。注意力机制本身就具有评估长序列全局依赖关系的能力 [20]。Dosovitskiy 等人 [33] 将纯 Transformer 模型应用于图像块，与基于 CNN 的方法相比取得了良好的效果。由于在任务相关的脑电图试验中，全局相互作用不可忽略，因此将 Transformer 模型引入脑电图处理。Kostas 等人 [34] 设计了一种使用 Transformer 模型进行脑电图分类任务的预训练和微调方法。Song 等人从空间和时间域进行特征学习，其中脑电图信号沿时间维度进行切片 [22]。刘等人 [35] 提出了类似的框架来处理脑电图的差分熵特征。Bagchi 等人 [23] 将脑电图转换为多帧活动图，然后使用基于 CNN 的模块以及结合 CNN 和 Transformer 的模块来捕获有用信息。然而，特征提取会减少原始数据中的信息量，并且往往依赖于特定任务。先前的研究通常侧重于如何提高脑电图解码的准确性，而忽略了从视觉上解读具有长期依赖关系的全局特征的作用。因此，受上述研究的启发，我们提出了脑电图一致性模型 (EEG Conformer)，将其作为一种高效的、具有新颖可视化方法的主干模型。

三、M

A. 概述

作为一种新兴的神经网络，Transformer 擅长捕捉全局依赖关系，但如何有效地将其应用于脑电信号解码仍有待探索。本文提出了一个称为 EEG Conformer 的全新框架，将 CNN 和 Transformer 直接结合起来，实现端到端的脑电信号分类。Conformer 借鉴 CNN 和 Transformer 的思想，利用卷积学习局部时空特征，然后采用自注意力机制来封装全局时空特征。

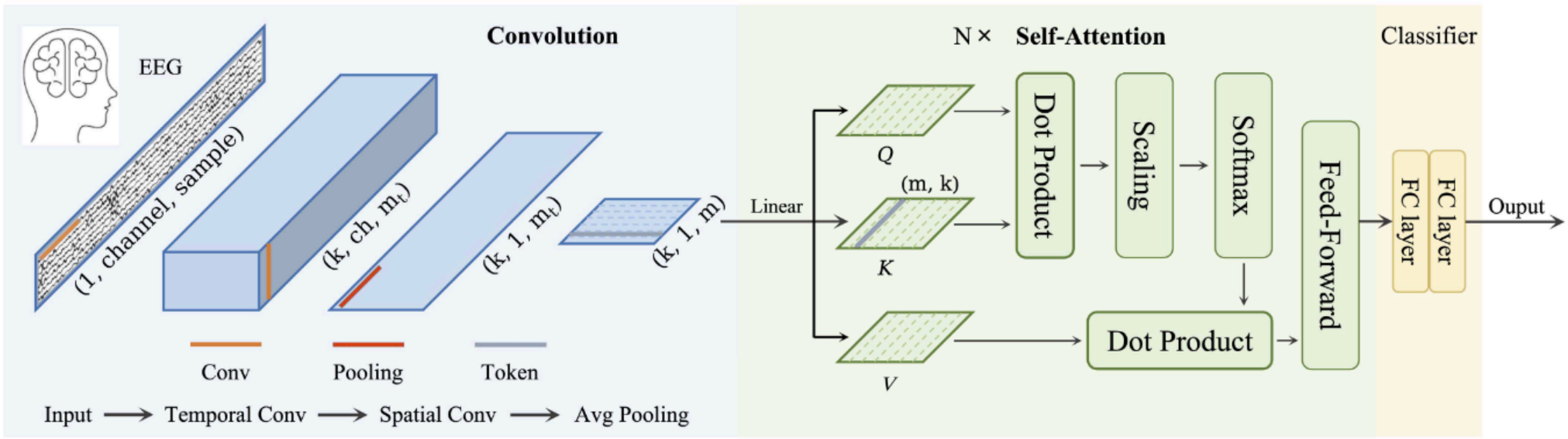

Fig. 1. The framework of Convolutional Transformer (Conformer), including a convolution module, a self-attention module, and a classifier module.

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,

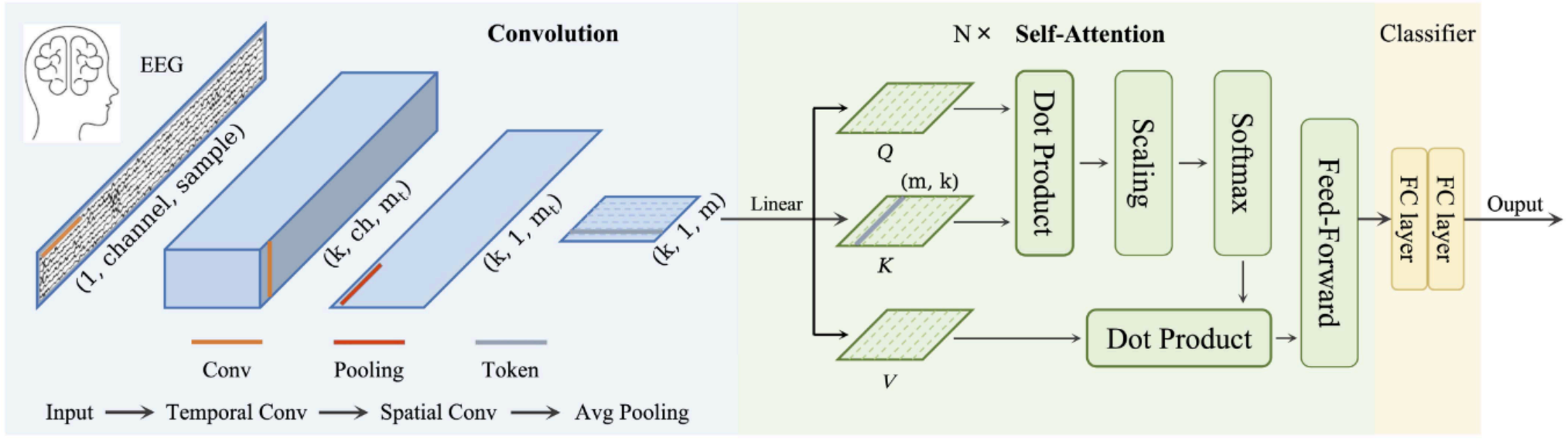

图 1. 卷积变换器（Conformer）的框架，包括卷积模块、自注意力模块和分类器模块。

总体框架如图 1 所示。该架构由三个部分组成：卷积模块、自注意力模块和全连接分类器。在卷积模块中，以原始二维脑电信号作为输入，分别沿时间维度和电极通道维度应用时间和空间卷积层。然后，利用平均池化层抑制噪声干扰并提高泛化能力。其次，将卷积模块获得的时空表示输入到自注意力模块。自注意力模块通过测量特征图中不同时间位置之间的全局相关性进一步提取长期时间特征。最后，采用由多个全连接层组成的紧凑分类器输出解码结果。

B.预处理

原始 EEG 试验的大小为 ch × sp，其中 ch 表示电极通道，sp 表示时间样本。无需引入额外的任务相关先验知识，我们仅使用几个步骤对原始脑电图数据进行预处理。首先，采用带通滤波滤除无关的高频和低频噪声。这里，我们使用 6 阶切比雪夫滤波器来保留与任务相关的节律。然后，进行 Z 分数标准化，以减少波动和非平稳性，如下所示

$$x = \frac{x - \mu}{\sqrt{\sigma}}, \tag{1}$$

其中 x 和 x 分别表示带通滤波数据和标准化输出。μ 和 σ 表示平均值和方差，用训练数据计算并直接用于测试数据。

C.网络架构

如图 1 所示，EEG Conformer 端到端流程由三个步骤组成：卷积模块、自注意力模块和全连接分类器。输入是一批经过预处理的脑电信号试验，其通道和样本维度相同，并扩展一维作为卷积通道。输出是不同脑电信号类别的概率。

表一
NA CM 架构

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）卷积模块：受[16]和[17]的启发，我们将二维卷积算子分成两个一维的时间和空间卷积层来设计卷积模块。第一层有 k 个大小为（1,25）的核，步长为（1,1），这意味着卷积在时间维度上进行。第二层保留 k 个大小为（ch，1）的核，步长为（1,1），其中 ch 等于 EEG 数据的电极通道数。该层充当空间滤波器，学习不同电极通道之间相互作用的表示。随后，采用批量归一化来增强训练过程并减轻过度拟合。我们遵循[17]使用指数线性单元（ELU）作为非线性的激活函数。第三层是沿时间维度的平均池化，核大小为（1,75），步长为（1,15）。该池化层平滑了时间特征，不仅避免了过拟合，还降低了计算复杂度。如表一所示，超参数 k 设置为 40。最后，我们重新排列了卷积模块的特征图，压缩了电极通道维度，并将卷积通道维度与时间维度进行转置。这样，我们将每个时间点的所有特征通道作为一个 token 输入到下一个模块中。

2）自注意力模块：我们假设上下文由于神经活动具有连贯性，低级时空特征中的依赖性表示将有利于脑电图解码。在此模块中，我们使用自注意力机制来学习脑电图特征的全局时间依赖性，以补充卷积模块中有限的感受野。来自上一个模块的有序标记被线性变换为等形的三元组，分别称为查询 (Q)、键 (K) 和值 (V)。对 Q 和 K 使用点积来评估不同标记之间的相关性。设计了一个缩放因子以避免梯度消失。

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

从而确保训练的稳定性。结果通过 So f tmax 函数得到加权矩阵，即注意力得分。然后，将注意力得分用点积加权到 V 上 [20]。该过程可以表示为

$$\text{注意力（Q，K，V）} = \text{So f tmax}\left(\frac{QK}{\sqrt{k}}\right)V, \tag{2}$$

其中 k 表示 token 的长度。此外，后面还连接了两个全连接前馈层，以增强拟合能力。此过程的输入和输出大小保持不变。整个注意力计算在自注意力模块中重复 N 次。

我们还采用了多头策略来进一步提升表征多样性。将 token 平均分成 h 个片段，分别输入自注意力模块，并将结果连接起来作为模块输出 [20]。该过程可以表示为

$$\begin{aligned}\text{MHA(Q, K, V)} &= [\text{头部}; \cdot\ \cdot\ \cdot ; \text{头部}],\\ \text{头部} &= \text{注意力（Q，K，V）}\end{aligned} \tag{3}$$

其中 MHA 代表多头注意力机制，Q、K、V∈R 分别表示第 l 个头中划分好的 token 通过线性变换得到的 query、key 和 value。

3）分类器模块：最后，我们采用两个全连接层作为分类器模块，经过 So f tmax 函数后输出一个 M 维向量。整个框架的损失函数采用交叉熵，公式如下：

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

其中 M 表示 EEG 类别的数量，y 和 $\hat{y}$ 分别为基本事实和预测标签。N 表示批次中的试验次数。

总结一下，首先将带通滤波和标准化的脑电数据输入模型。然后，数据依次经过时间和空间卷积层，并由池化层整理成 token。之后，使用 N 个自注意力层，再经过全连接层输出分类结果。

IV. 急诊室

在本节中，我们将在三个公开的脑电图数据集（包括流行的运动想象和情绪识别范式）上进行实验，以验证所提出的网络。我们不仅将我们的方法与不同的先进方法进行了比较，而且还通过消融研究展示了引入基于注意力机制的 Transformer 所带来的改进。我们还提供了详细的对比实验，以展示注意力参数对整体性能的影响。最后，我们设计了不同的可视化方法以提高可解释性。

A.数据集

我们在三个广泛使用的 EEG 数据集上评估了我们的方法，包括 BCI 竞赛 IV 数据集 2a、BCI 竞赛 IV 数据集 2b，SEED [36] 这些 EEG 数据集是使用不同的采集设备、范例、受试者数量和样本量收集的，从而公平地验证了我们方法的泛化能力。

1）数据集 I：BCI 竞赛 IV 数据集 2a 由格拉茨技术大学提供，包含 9 名受试者的脑电数据。数据集包含四项运动想象任务，涵盖想象移动左手、右手、双脚和舌头的动作。数据采集于不同日期的两个时间段，使用 22 个 Ag/AgCl 电极以 250 Hz 的采样率进行。一个时间段包含 288 个脑电图试次，即每个任务 72 个试次。我们在每个试次中使用 [2, 6] 秒，并使用带通滤波器将脑电图数据滤波至 [4, 40] Hz，如同实验中的 [8] 所示。第一个时间段用于训练，第二个时间段用于测试。

2）数据集 II：BCI 竞赛 IV 数据集 2b 由格拉茨技术大学提供，包含 9 名受试者的脑电数据。数据集包含两个运动想象任务，分别涵盖左右手运动的想象。数据集使用三个双极电极（C3、Cz 和 C4）采集了五个会话，采样率为 250 Hz，每个会话包含 120 个试次。实验中使用了每个试次的 [3, 7] 秒。此外，我们还在 [4, 40] Hz 之间进行了带通滤波，以降低高频和低频噪声。前三个会话为训练集，后两个会话为测试集。

3）数据集三：上海交通大学提供的 SEED 数据集包含来自 15 位受试者的情绪脑电信号。这些信号由 15 个电影片段刺激，分别对应积极、中性和消极三种情绪。数据采集过程对每位受试者重复三次，间隔约一周。脑电信号由 62 个电极采集，采样率为 1000 Hz，随后下采样至 200 Hz。每个样本以不重叠的一秒时间窗口进行分割，共计 3394 次试验。我们还对数据进行了 [4, 47] Hz 的带通滤波。SEED 数据集采用了五折交叉验证。

B.数据增强

脑电图采集耗时较长，导致数据集过小，容易出现过拟合。一些方法采用数据增强的方法为模型提供足够的样本 [16]。然而，传统的添加高斯噪声或裁剪等策略可能会进一步降低信噪比或破坏原有的相干性。因此，我们采用时间域的分割与重建 (S&R) 来生成新的数据。借鉴 [37] 的方法，将同一类别的训练样本均等地分成 N 个段，然后在保持原始时间顺序的情况下随机连接起来。我们在每次迭代中生成与批次大小相同的增强数据。

C. 实验细节

我们的方法基于 Python 3.10 中的 PyTorch 库和 Geforce 3090 GPU 实现。我们使用 2 个

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

https://www.bbci.de/competition/iv/desc_2b.pdf 训练模型。

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.

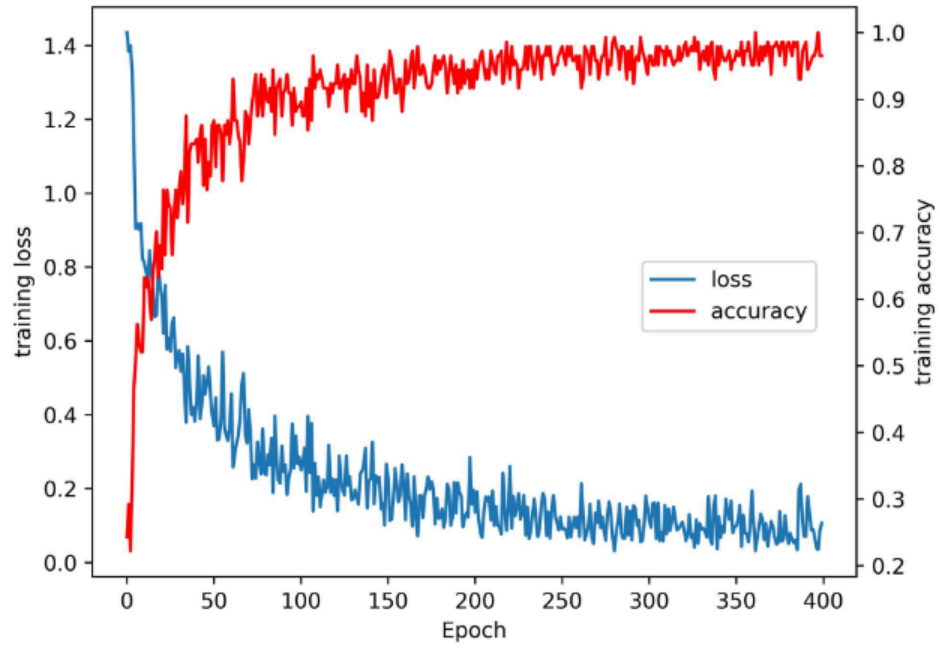

Fig. 2. Loss and accuracy during training of EEG Conformer.

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

Adam 优化器，学习率 β 和 β 分别为 0.0002、0.5 和 0.999。我们将自注意力执行次数 N 设置为 6，头节点数 h 设置为 10，S&R 中的 N 设置为 8。分类准确率和 kappa 值作为整体性能的评估指标。kappa 值可以通过以下公式计算：

$$\text{卡帕值} = \frac{p - p}{1 - p}, \tag{5}$$

其中，p 表示所有试验的平均准确率，p 表示随机猜测的准确率。采用 Wilcoxon 符号秩检验（Wilcoxon SignedRank Test）来分析统计显著性。

D. 基线比较

我们进行了广泛的基于主题的实验，并在三个公共数据集上将我们的方法与一些最先进的方法进行了比较。

数据集 I 是目前使用最广泛的多类运动想象数据集。我们比较了许多代表性方法，这些方法在该数据集上取得了令人印象深刻的性能。例如，使用手工制作的空间特征的 BCI 竞赛 IV 的获胜者 FBCSP [8]；ConvNet [16] 和 EEGNet [17]，它们在基于 CNN 的端到端框架下在许多 EEG 数据集上取得了显著的效果；C2CM [28] 将 FBCSP 特征输入到 CNN 模型中，结合了传统特征提取和深度学习方法的优势；FBCNet [38] 通过空间滤波多视图数据来提取光谱空间特征。我们甚至与基于深度表示的域自适应 (DRDA) [39] 进行了比较，DRDA 利用来自其他受试者的数据通过对抗学习进行增强。

表 II 显示了每个受试者的分类性能和在数据集 I 上的平均结果。我们可以观察到，与依赖于传统特征提取的 FBCSP 相比，我们的 Conformer 将准确率提高了 10.91%（$p < 0.01$）。结果还表明，其他深度学习方法（例如 ConvNet 和 EEGNet）优于 FBCSP，这表明基于 CNN 的方法具有强大的特征表示能力。然而，这些基于 CNN 的方法由于感知场有限而仅关注局部特征，而忽略了全局相关性，这可能会损害相干 EEG 序列的解码准确性。不同的是，我们的方法在原始 CNN 的基础上集成 Transformer 架构，封装了局部和全局依赖关系。因此，Conformer 在大多数受试者上获得了更好的结果，并在平均准确率（$p < 0.05$）和 kappa 上实现了显着的提升。 C2CM 和 FBCNet 有效地结合了手工特征和深度模型的思想，但除了主题 5 之外，仍然无法超越我们的成果（$p < 0.05$），尽管 C2CM 针对每个主题微调了模型参数。DRDA 引入了分布与目标主题一致的其他主题的数据，其效果仍然不如我们仅使用目标主题数据的方法（$p < 0.05$），这再次证明了同时利用局部和全局特征的有效性。

然后，我们在表三中展示了在数据集 II 上与几种最先进的方法的比较。我们可以看到，二分类结果呈现出与数据集 I 相似的趋势。

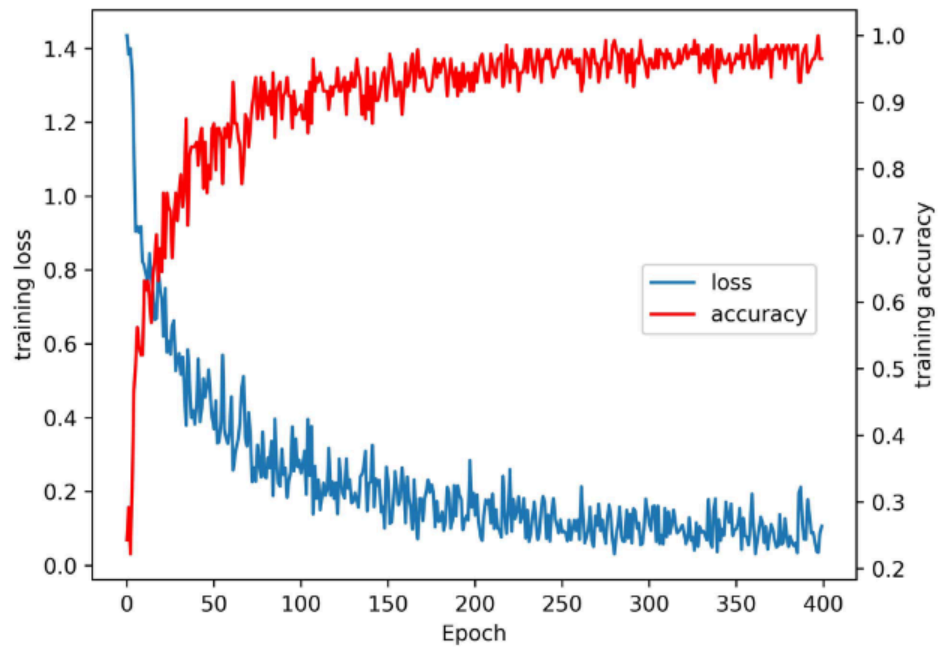

图 2.EEG Conformer 训练期间的损失和准确度。

Conformer 的整体性能相比 FBCSP 有显著提升（$p < 0.05$），在样本 1 上准确率甚至提升了 12.5%。与其他仅使用 CNN 架构的端到端方法相比，ConvNet（$p < 0.05$）和 EEGNet（$p < 0.01$）的提升更为显著，分别为 5.25% 和 4.15%。在几乎所有样本上，我们方法的平均准确率和 kappa 值仍然优于 DRDA，这进一步验证了我们方法的有效性。

我们还在多类别脑电情绪数据数据集 III 上全面评估了我们的方法。我们将这些方法与以下机器学习方法进行了比较：SVM [36]，它们首先在该数据集上取得了显著的效果；图正则化极限学习机 (GELM) [40]，它具有单个前馈层来学习判别特征；以及区域到全局时空神经网络 (R2G-STNN) [42]，它采用双向长短期记忆来学习情绪脑电信号的时空特征。此外，我们还将学习不同脑电通道之间内在关系的基于图的神经网络（例如动态图卷积神经网络 (DGCNN) [41] 和正则化图神经网络 (RGNN) [43]）进行比较。结果列于表 IV 中。可以看出，与其他最先进的方法相比，Conformer 在数据集 III 上仍然具有竞争力。通过这种方式，我们的方法在运动想象和情绪识别范式上都取得了令人印象深刻的性能，说明了我们的方法具有良好的泛化能力。

E. 训练过程

在图像处理中，Transformer 模型通常需要大量数据进行预训练才能在下游任务中取得良好的效果。然而，EEG Conformer 模型由于校准数据有限，并未进行预训练。我们在图 2 中展示了训练过程中损失和准确率的变化趋势。在轻量级自注意力模块的帮助下，该过程保持稳定。可以注意到，模型在 250 个 epoch 左右快速收敛。此外，我们的方法也很有效。我们使用数据集 I 中的第一个受试者对 Conformer 模型进行了 2000 个 epoch 的连续训练。

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.

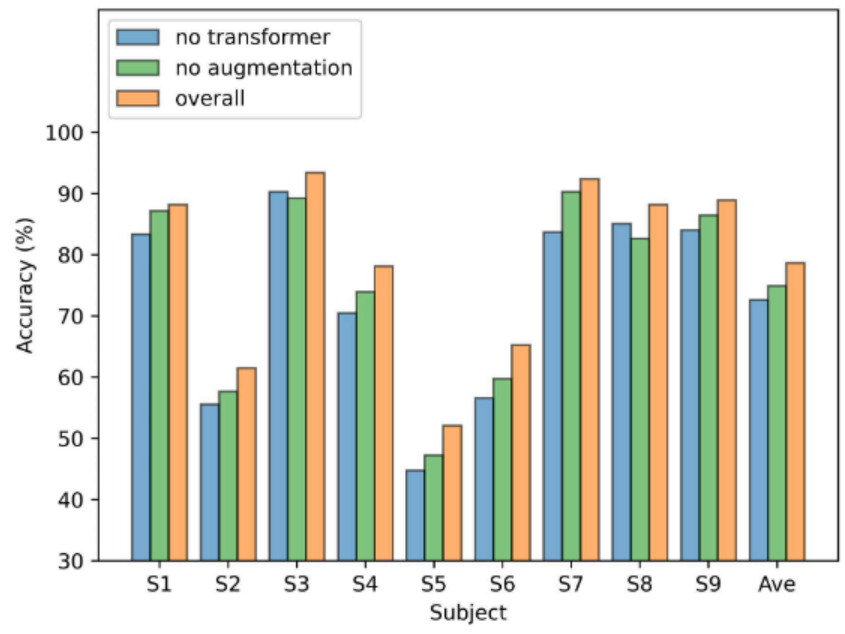

Fig. 3. Ablation study on the self-attention module and data augmentation.

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

表二

CWS---AMDI

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**

表三

CWS - - -AMD 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**

表 IV CWS - - -AMD 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**

在单个 GPU 上，每个 epoch 的平均时间为 0.27 秒。

F. 消融研究

EEG Conformer 相较于基于 CNN 的方法，其关键改进在于增加了基于注意力机制的 Transformer 模块，用于学习全局表征。此外，数据增强也可能对最终结果有所贡献。因此，我们对数据集 I 进行了消融研究，如图 3 所示，分别移除了自注意力模块和 S&R 数据增强。可以看出，移除 Transformer 部分后，每个受试者的结果都有显著下降。受试者 6 下降最多，为 8.68%，受试者 3 下降最少，为 3.12%。平均准确率显著下降了 6.02%（$p < 0.01$）。与 ConvNet [16] 类似，图 3 中的实验结果也表明数据增强策略有助于提升我们模型的性能。与没有数据增强相比，整体性能平均提高了 3.75%（$p < 0.01$）。有趣的是，对于区分度较好的受试者 1，提升幅度仅为 1.04%，而对于原本表现较差的受试者 5 和 6，提升幅度更为显著，分别达到了 4.86%和 5.56%。因此，在训练过程中引入数据增强技术，增强了 Conformer 的鲁棒性。

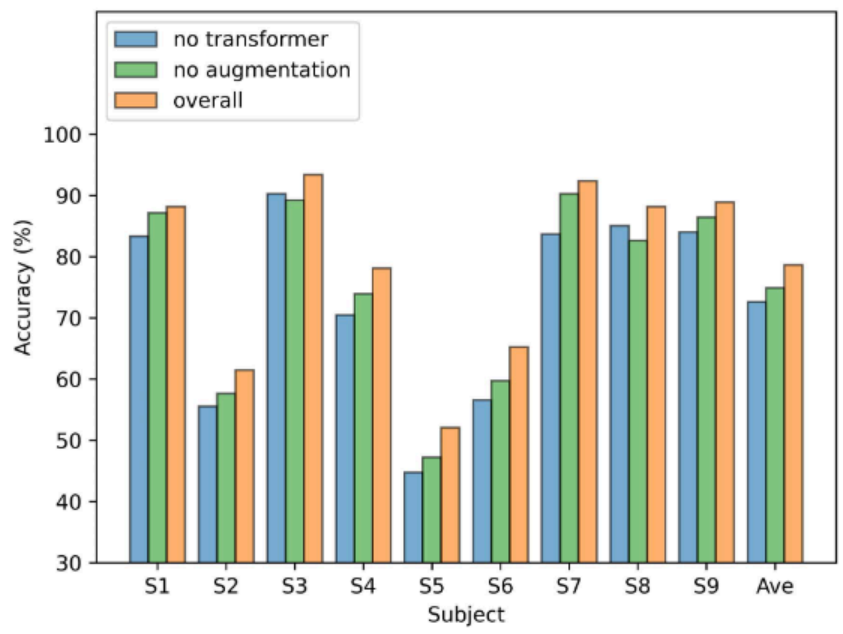

图 3. 自注意力模块和数据增强的消融研究。

G. 参数灵敏度

在本节中，我们详细评估了自注意力模块中几个重要参数对性能的影响。这些参数包括自注意力层的深度 N、注意力头的数量 h，以及池化核的设计（池化核用于构建用于学习全局特征的输入）。

深度通常是影响端到端模型（如 CNN、Transformer）拟合能力的关键因素。如图 4 所示，我们通过逐步增加自注意力模块的层数（从 0 到 15）来探索深度对 EEG Conformer 的影响。可以看出，对于数据集 I，当深度从 0 增加到 1 时，准确率有显著的提高（$p < 0.01$），这再次说明了 Transformer 的引入确实有助于 EEG 解码。对于其他深度，最高精度仅比最低精度高 1.24%，且差异不显著（$p > 0.05$）。然而，如参数曲线所示，数量

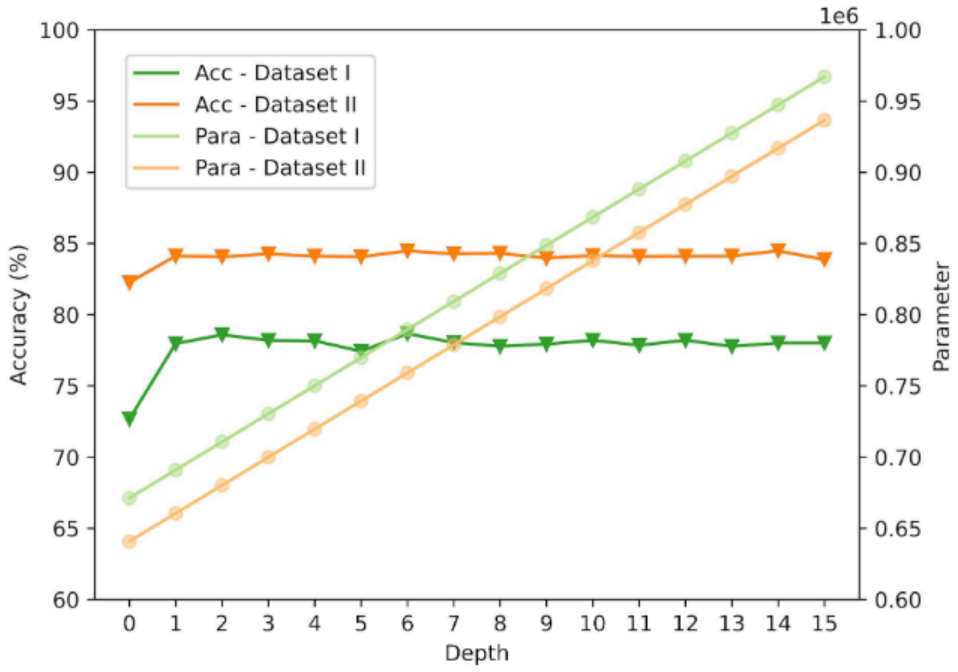

Fig. 4. The influence of the depth of the self-attention module (from 0 to 15) on the accuracy and the amounts of parameters for Dataset I and II.

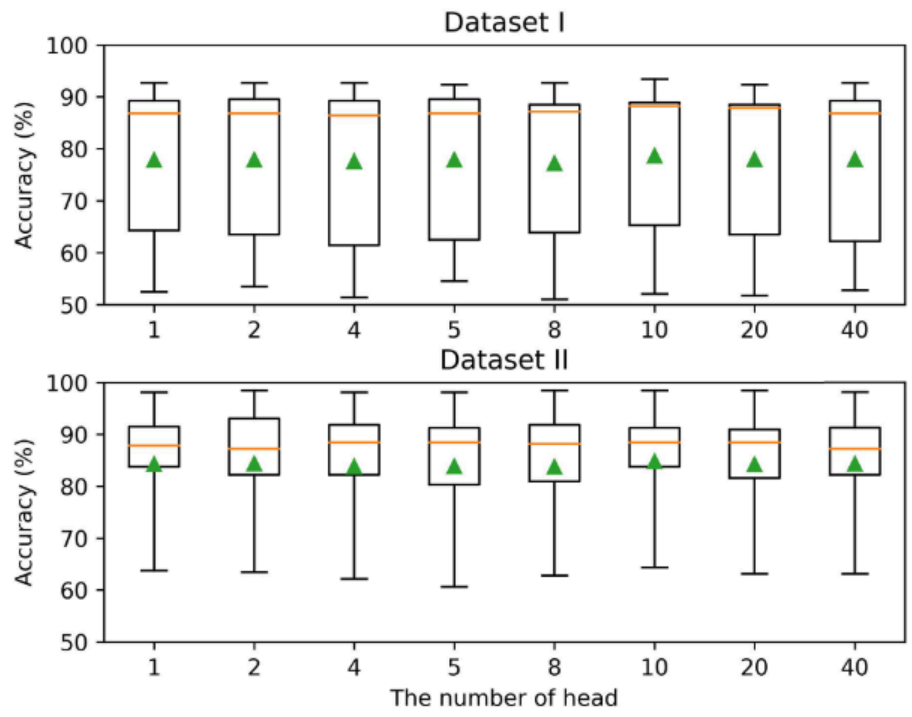

Fig. 5. The influence of the number of attention heads on the accuracy for different datasets.

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.

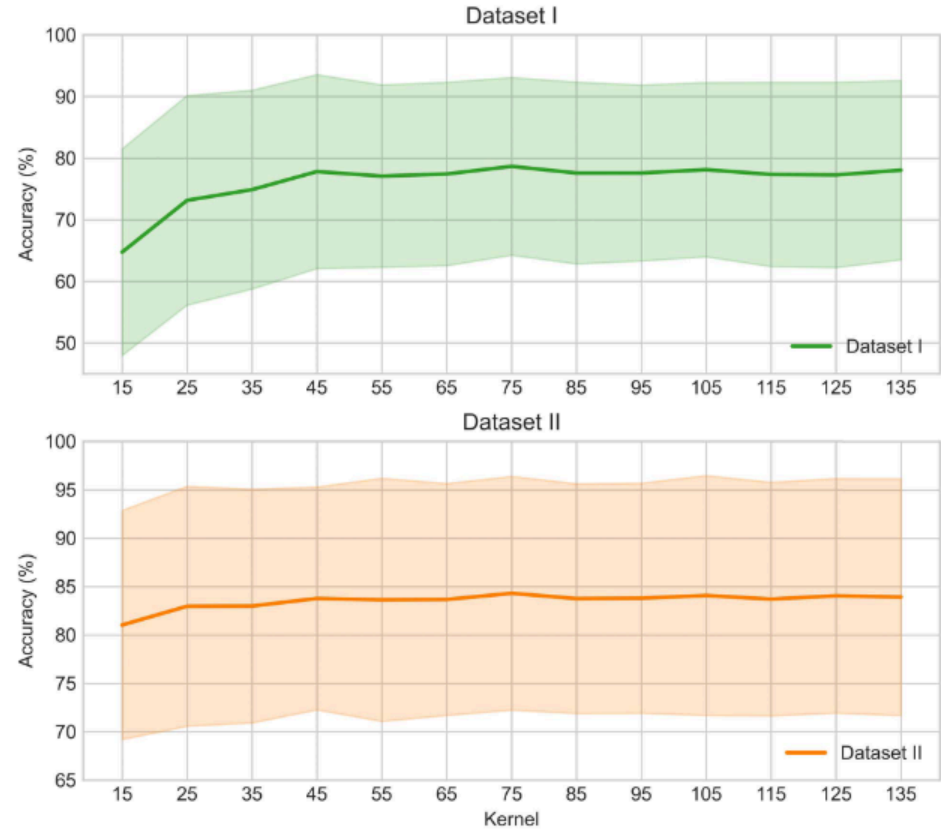

Fig. 6. The influence of different kernel sizes in the pooling layer, namely, the token size of the self-attention module.

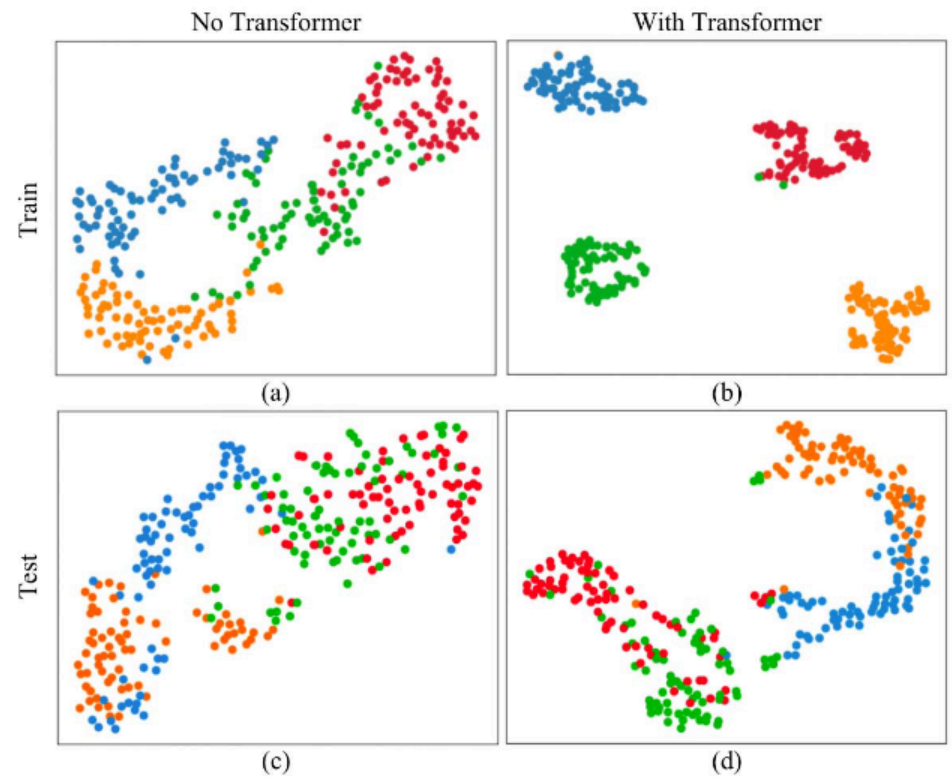

Fig. 7. t-SNE visualization illustrates the significance of introducing Transformer for feature learning. Different colors represent different categories.

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.

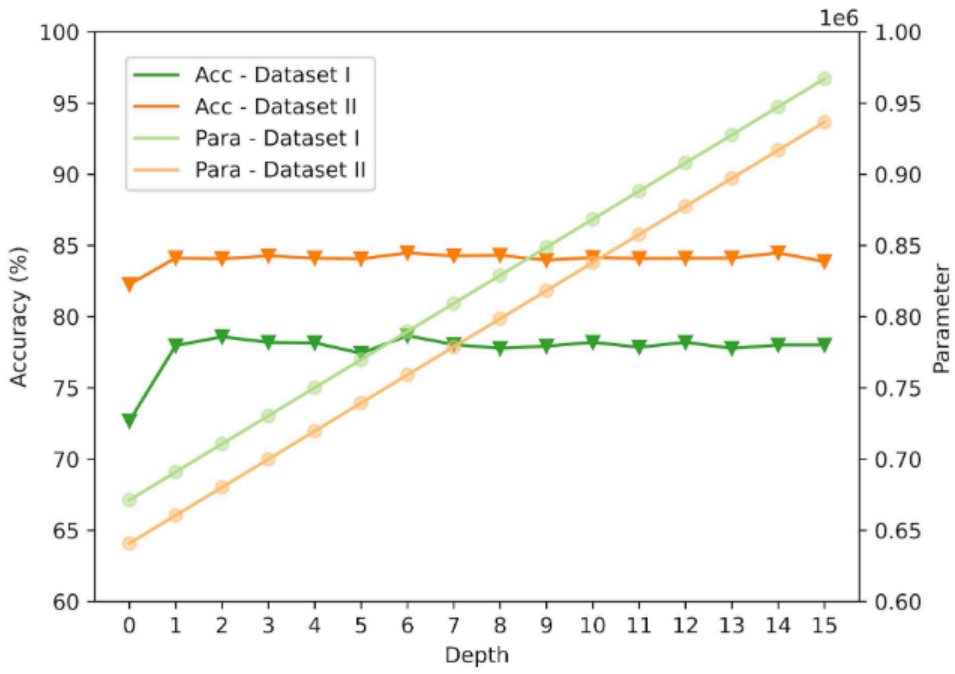

图 4. 自注意力模块深度（从 0 到 15）对数据集 I 和 II 的准确率和参数数量的影响。

图 5. 注意力头数量对不同数据集准确率的影响。

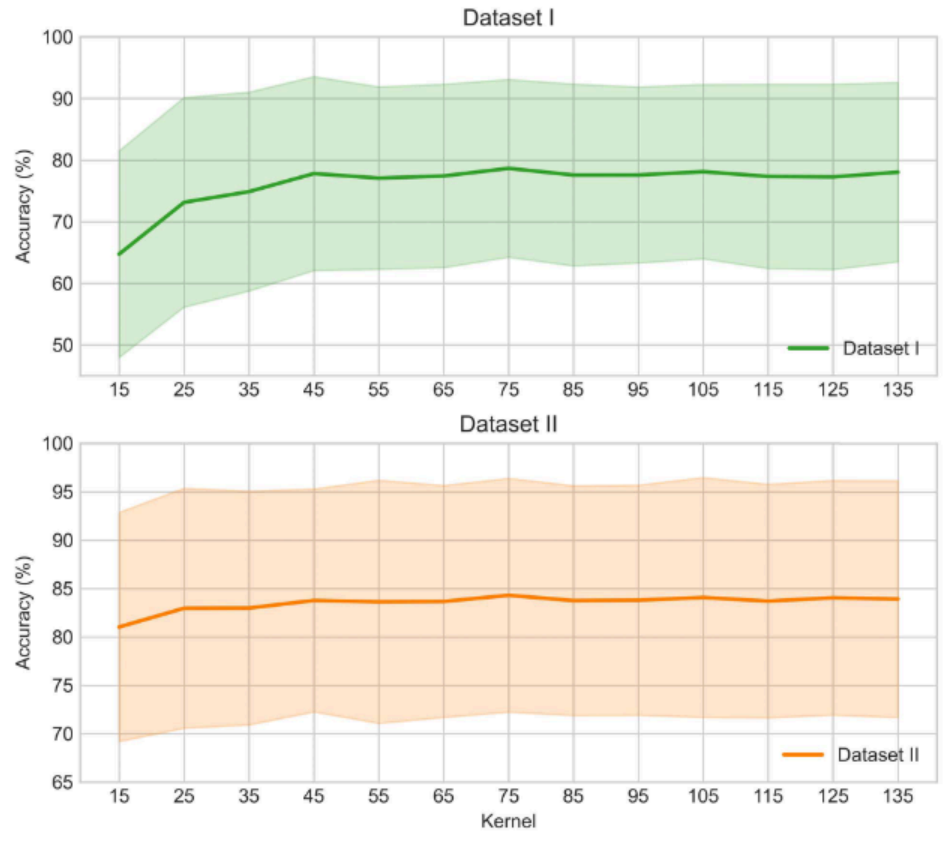

图 6. 池化层中不同核大小的影响，即自注意力模块的 token 大小。

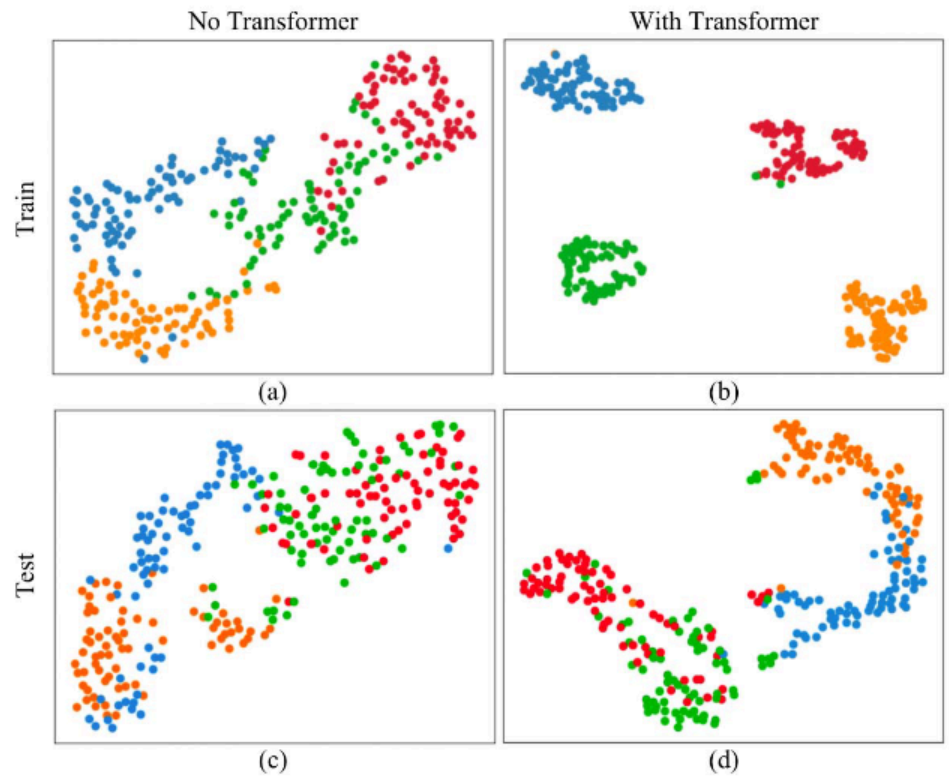

图 7. t-SNE 可视化展示了引入 Transformer 对特征学习的重要意义。不同颜色代表不同的类别。

参数数量随着深度成正比增加，这使得模型的性价比较低。在数据集 II 上的相同评估也显示出 Conformer 对 selfattention 深度的不敏感性。

头是常见的基于多头注意力机制的 Transformer 模型中的重要参数，据报道它可以帮助学习特征的不同方面。我们也比较了不同头选择对模型的影响，如图 5 所示，选择了 1 到 40 之间的 8 个头编号。从框中看，不同头编号对结果的影响没有明显的规律，不同受试者的分布没有明显差异。平均准确率保持着平缓的波动，在数据集 I 上的波动范围仅为 1.43%，在数据集 II 上的波动范围仅为 1.02%。随着头数量的增加，性能略有上升趋势，但随后有所下降。当头编号取 10 时，数据集 I 的平均准确率比头编号取 1 时高 0.82%，数据集 II 的平均准确率高 0.50%（$p>0.05$）。总体而言，改变头编号尚未表现出对特征学习的显著促进作用。

池化核所确定的 token 也是自注意力模块的关键因素。如果核太大，时间特征会过于平滑，丢失有用的细节，模型很难感知细节之间的全局关系。相反，如果核太小，性能容易受到局部噪声的影响。我们比较了不同池化核选择对模型性能的影响，如图 6 所示。核大小从 15 到 135，间隔为 10。可以清楚地看到，当核大小开始增加时，平均准确率有显著的提升。在数据集 I 上，通过将核从 15 增加到 45，获得了 13.08%的增益（$p<0.01$）。此后，结果趋于平缓，随着核大小的增加，并没有明显上升。实验表明，对于信噪比较低的 EEG，将自注意力应用于足够大的切片确实有意义。

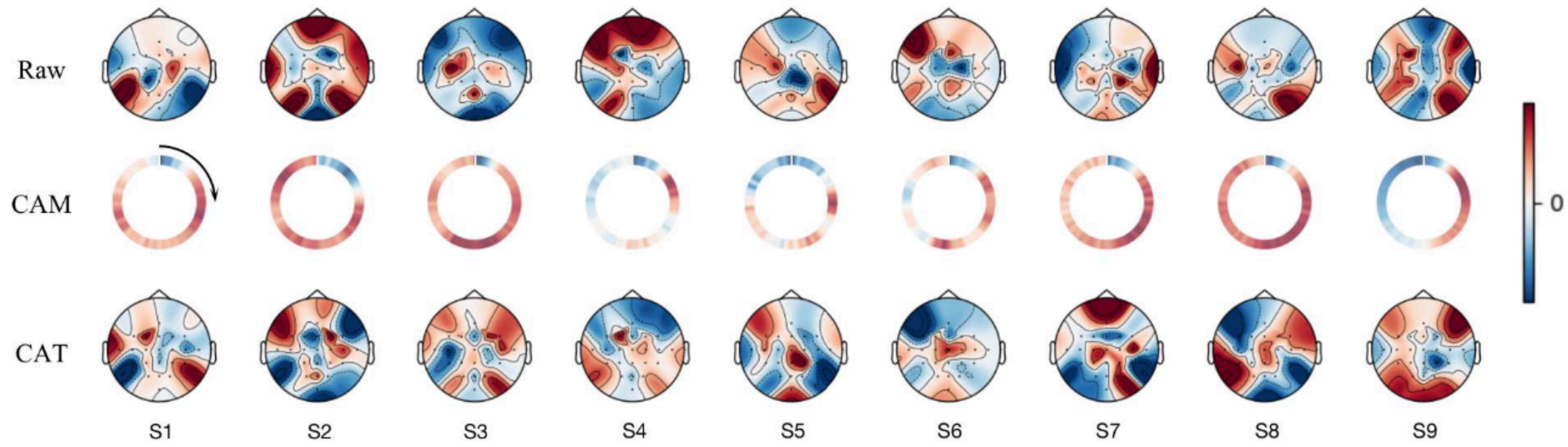

Fig. 8. Raw EEG topography averaged over all trials of each subject, Class Activation Mapping (CAM) of the Transformer module on the input EEG, Class Activation Topography (CAT) we designed to show CAM-weighted EEG. Raw shows that many regions are activated throughout the trial. CAM shows that our model 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.

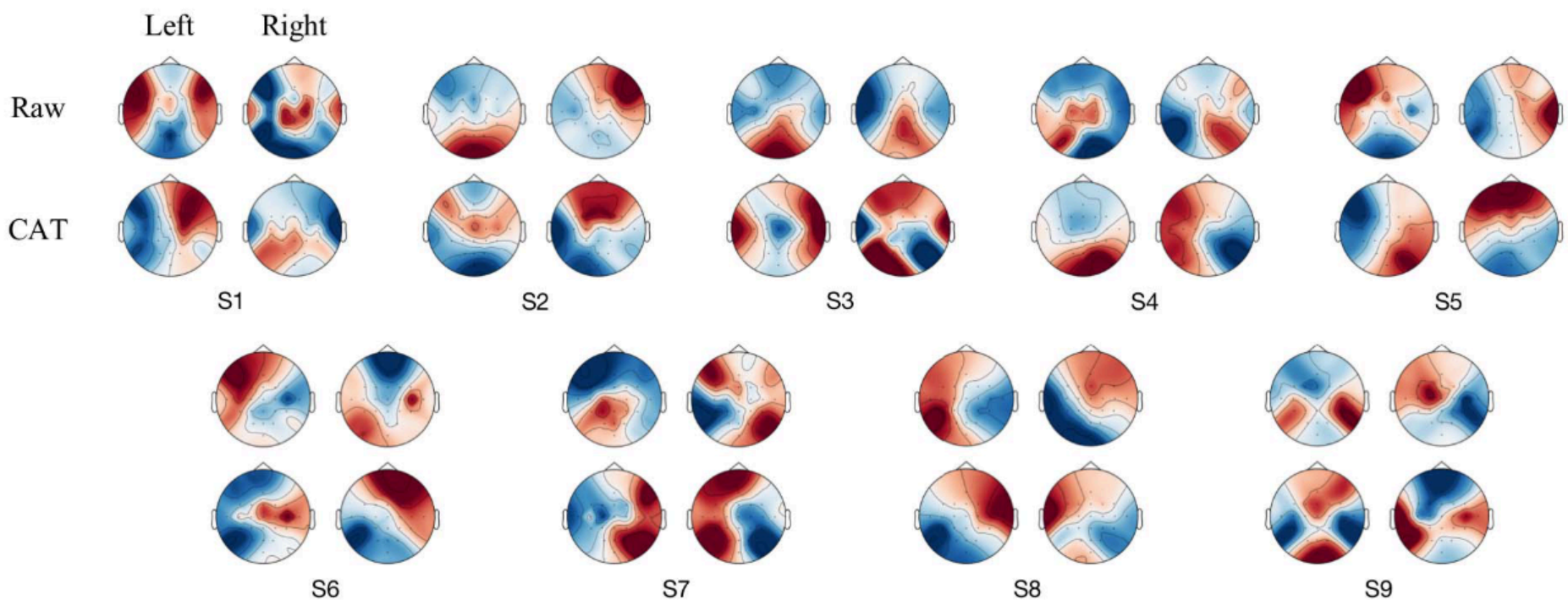

Fig. 9. CAT shows the ERD/ERS phenomena on both the data of imagining left and right hand movements, compared to the irregular patterns in raw EEG topography. Contralateral activation and ipsilateral inhibition can be clearly observed in the CAT of several subjects, such as S1, S7, and S8.

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.

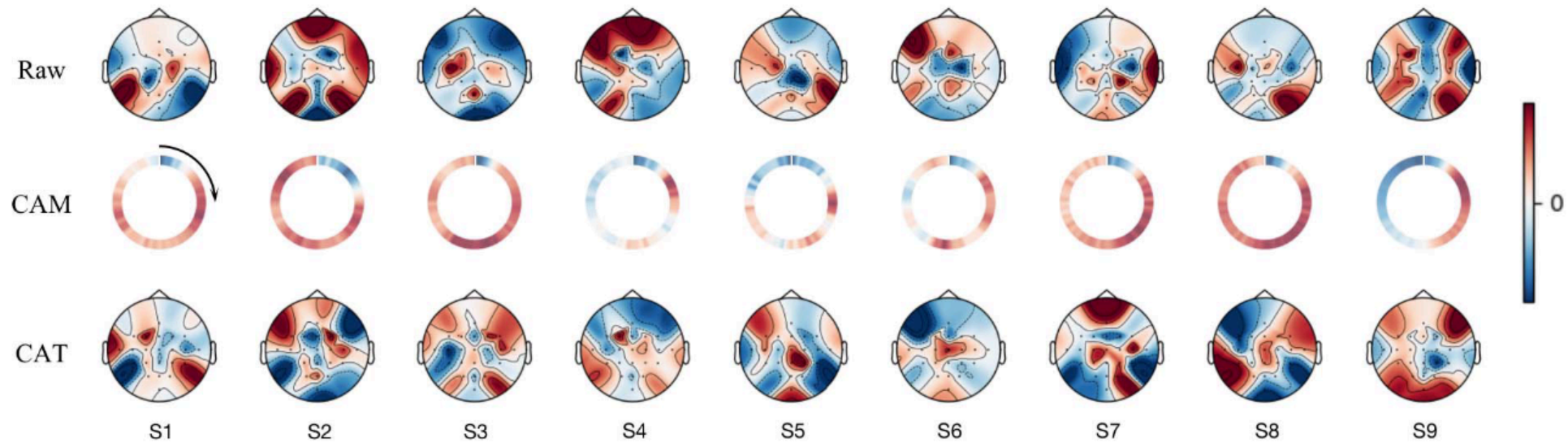

图 8. 原始脑电图拓扑图，取每个受试者所有试验的平均值，Transformer 模块在输入脑电图上的类激活映射 (CAM)，以及我们设计的类激活拓扑图 (CAT)，用于显示 CAM 加权脑电图。原始数据表明，在整个试验过程中，许多区域都被激活。CAM 表明我们的模型对时域中的不同范围给予了不同的关注。CAT 表明我们的模型关注运动图像数据中运动皮层的区域。

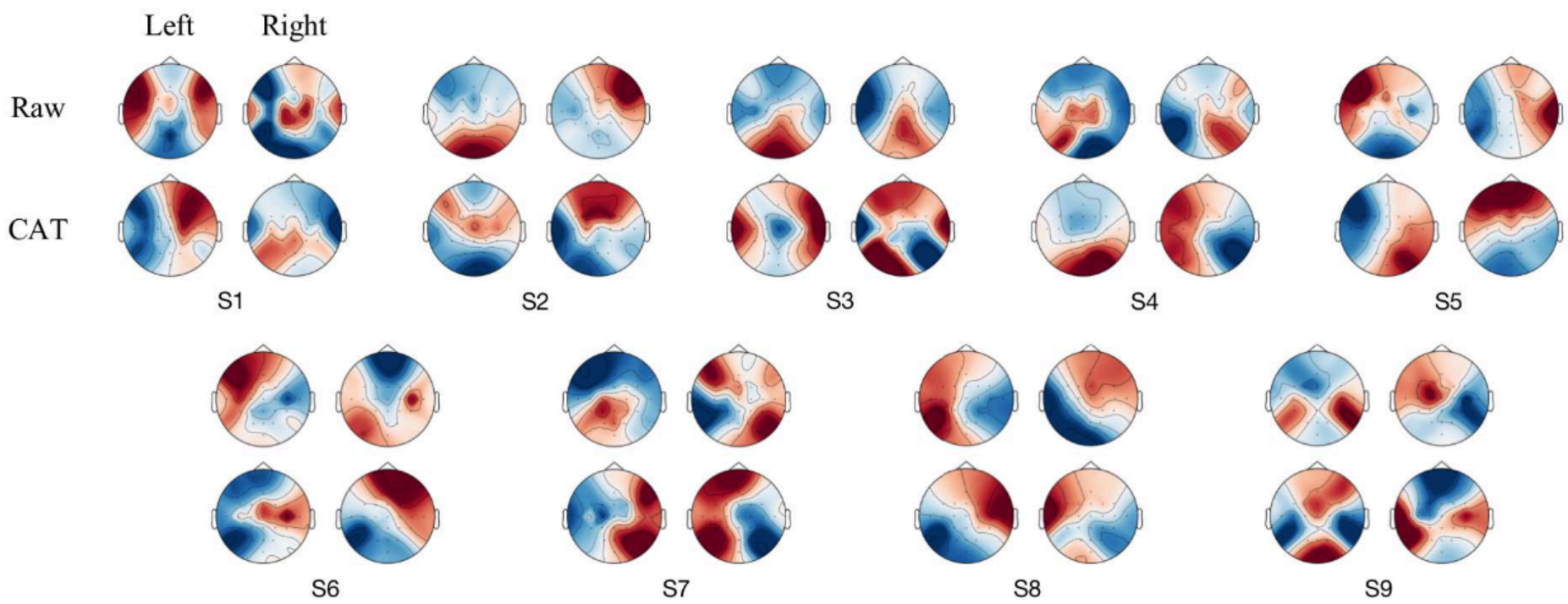

图 9. CAT 显示了想象左手和右手运动数据上的 ERD/ERS 现象，与原始 EEG 拓扑图中的不规则模式相比。在 S1、S7 和 S8 等多个受试者的 CAT 中可以清晰地观察到对侧激活和同侧抑制。

H. 可视化

我们从两个角度可视化地展示 EEG Conformer 的可解释性，包括 t-SNE [44] 的深度特征和地形上反映的时空特征。

1）特征分布：t 分布随机邻域嵌入（t-SNE）是一种流行的统计降维和可视化方法。图 7 展示了数据集 I 中受试者 1 在加入和未加入 Transformer 充分训练后的特征分布。我们可以看到，对于训练数据，在没有 Transformer 的情况下，不同类别的特征比较接近。加入 Transformer 后，类别间距离变大，类别内距离变小，如图 7(b)所示。另一方面，在没有 Transformer 的情况下，类别间混叠现象较为明显，使得类别边界更加清晰，如图 7(d)所示。

2) 全局表征：引入 Transformer 来学习脑电图数据中的全局时间依赖性，这意味着从时间序列中定位更重要的信息，以完成解码任务。我们使用拓扑结构和梯度加权类激活映射 (CAM) [45] 来展示我们的模型在运动想象数据集 I 上学习到的全局表征，如图 8 所示。图中第一行表示对每个受试者的所有训练试验的拓扑结构取平均值。不同受试者之间没有明显的活跃脑区线索。采用 CAM 来监测自我注意模块在 EEG 特征上关注的时间段，如图 8 第二行所示。在运动想象过程中，EEG 数据被绘制成一个圆圈，从顶部顺时针旋转。不同的激活在不同的时间呈现。正如预期的那样，所有受试者的数据在试次开始时都呈衰减状态，这可能表明运动意图的潜伏期存在。

我们进一步提出了一种新的脑电图可视化方法，即类激活拓扑图（CAT）。脑电图绘制在乘以归一化 CAM 后的归一化数据上。从图 8 第三行开始，经 CAM 加权后的脑电图数据大部分集中在运动皮层区域，这与运动想象范式[46]相一致。此外，图 9 分别绘制了想象左手运动和右手运动的原始脑电图和 CAT 图。我们惊讶地发现了事件相关去同步（ERD）和事件相关同步（ERS）现象。与不规则的原始脑电图拓扑图相比，在第 1 例和第 8 例等几例受试者的 CAT 图中观察到明显的对侧激活和同侧抑制。

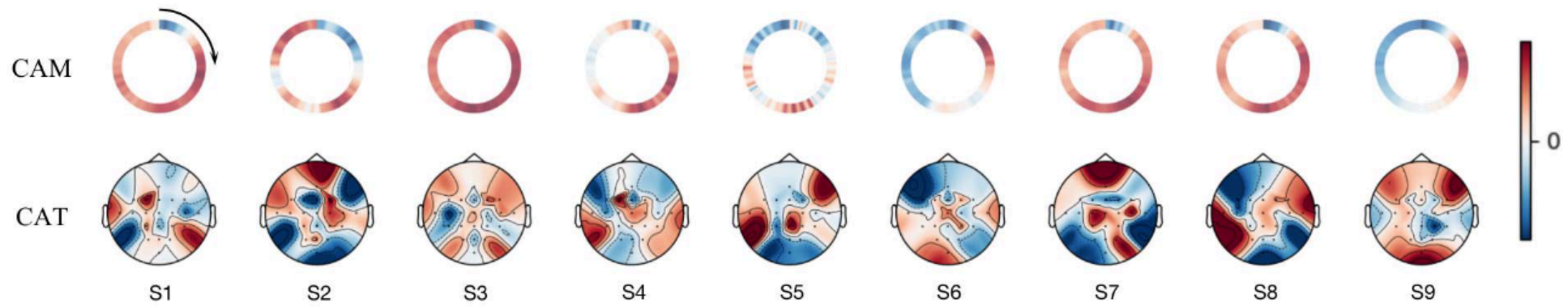

Fig. 10. CAM and CAT of the model with only 1 head in the self-attention module. The activation is close to that in Fig. 8 with 10 heads.

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.

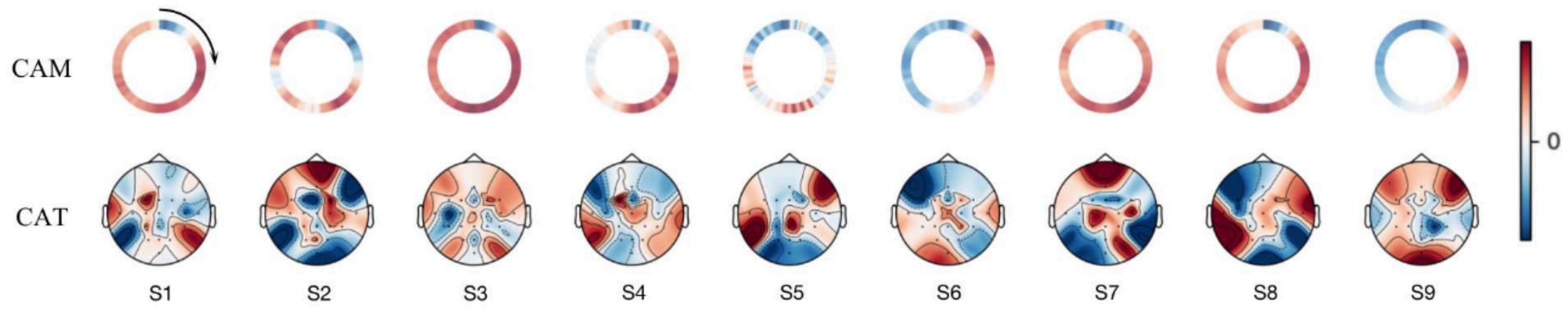

图 10. 自注意力模块中仅有 1 个注意力头的模型的 CAM 和 CAT。激活函数与图 8 中 10 个注意力头的模型接近。

V. D

脑机接口 (BCI) 系统的实用性取决于解码方法的性能。我们提出了一种名为 Conformer 的简洁有效的方法，该方法结合了 CNN 和 Transformer 网络的优势。Conformer 是一种无需预训练的轻量级脑电信号解码解决方案。它仅采用几个预处理步骤，包括带通滤波和标准化，且不太依赖于特定任务。同时具有时间和空间卷积层的卷积模块关注低级表示，考虑局部时间特征；而自注意力模块则进一步关注长期依赖关系并捕捉全局时间相关性。因此，与现有的基于 CNN 的模型相比，该方法能够学习更具判别性的表示。

实验结果表明，EEG Conformer 在三个采用不同范式和数据采集方式的数据集上取得了最佳结果。消融研究表明，Transformer 模块对整体模型贡献显著，数据增强有助于提升训练效果。我们还探讨了几个关键参数对模型的影响。结果表明，该模型对自注意力模块的深度和头单元数量不敏感。然而，池化层的核大小表现出显著的影响，这表明使用较大的单元来施加注意力有助于避免局部噪声的干扰。详细的可视化效果用于说明模型的可解释性。Transformer 模块与 t-SNE 所示的特征分布相比，具有更好的判别能力。我们还设计了一种新的可视化方法 CAT，结合脑电图拓扑结构和运动图像增强 (CAM) 来发现模型中某一层的功能。结果表明，我们的模型能够利用运动意象数据，聚焦于运动皮层附近的变化。此外，左右手意象产生的运动皮层变化 (ERD/ERS) 也能清晰地感知到。

由于多头在自注意力模块中的作用尚不明确，因此我们针对数据集 I 训练了仅使用 1 个头的模型，并在图 10 中绘制了 CAM 和 CAT。我们可以看到，自注意力模块的激活值与图 8 中 10 个头的激活值接近。对比表明，两种情况都学习到了相似的全局特征，从而获得了相似的解码准确率。激活值上的细微差异仍有待解决。

还有一些局限性。首先，我们主要验证振荡脑电数据，例如运动想象和情绪，它们缺乏像事件相关电位 (ERP) 脑电数据那样的平稳模式。其次，当前模型的规模并不小。对于数据集 I，与移除自注意力模块相比，Conformer 的参数增加了 17.6%。这些额外的成本来自于用于计算全局依赖关系的线性变换和前馈层。虽然我们已经确认训练模型的时间成本在实际使用中是可以接受的，但这仍然是一个可以改进的问题。此外，全连接分类器贡献了大量的参数。全局平均池化可以作为替代方案，且性能损失较小。第三，所提出的方法是基于每个人进行训练和验证的，无法利用来自其他受试者的有用信息。我们将在未来将该模型应用于 ERP 和与受试者无关的任务中。

VI.C

本文提出了一种简洁高效的脑电信号解码方法——Conformer。该方法将 Transformer 与 CNN 相结合，用于学习时域的全局依赖关系。通过详细的对比实验，我们在不同的脑电信号数据集上取得了显著的效果。可视化结果还表明，我们的模型能够在全局层面找到符合范式原则的关键信息。总体而言，我们的模型在提升脑电信号解码方面取得了良好的效果。

R

[1] J. Jin 等，“一种利用脑电图图形表示进行基于运动想象的脑功能评估的新型分类框架——
计算机接口”，IEEE Trans. Neural Syst. Rehabil. Eng.，第 30 卷，第 20-29 页，2022 年。
[2] B. Liu、X. Chen、N. Shi、Y. Wang、S. Gao 和 X. Gao，“通过任务判别提高单独校准的 SSVEP-BCI 的性能
组件分析”，IEEE Trans. Neural Syst. Rehabil. Eng.，第 29 卷，第 1998-2007 页，2021 年。
[3] JW Li 等人，“利用脑节律测序实现基于脑电图的情绪识别的单通道选择”，IEEE J. Biomed. Health Informat.，第 26 卷，第 6 期，第 2493-2503 页，2022 年 6 月。[4] S. He 等人，“基于脑电图和眼电图的异步混合脑机接口：集成拼写器、网页浏览器、电子邮件客户端和文件浏览器的系统”，
IEEE Trans. Neural Syst. Rehabil. Eng.，第 28 卷，第 2 期，第 519-530 页，2020 年 2 月。
[5] K.-T. Kim、H.-I. Suk 和 S.-W. Lee，“利用稳态体感诱发电位控制脑控轮椅”，IEEE
Trans. Neural Syst. Rehabil. Eng.，第 26 卷，第 3 期，第 654-665 页，2018 年 3 月。
[6] Y. Song、S. Cai、L. Yang、G. Li、W. Wu 和 L. Xie，“一种基于 EEG 的实用人机界面，用于在线控制上肢辅助机器人，”Frontiers Neurorobot.，第 14 卷，第 32 页，2020 年 7 月。[7] B.-Y. Tsai、SVS Diddi、L.-W. Ko、S.-J. Wang、C.-Y. Chang 和 T.-P. Jung，“基于赫布/反赫布学习网络的自适应伪影子空间重建开发，用于增强 BCI
性能”，IEEE 神经网络学习系统汇刊，抢先体验版，2022 年 6 月 17 日，doi：10.1109/TNNLS.2022.3174528。
[8] KK Ang、ZY Chin、C. Wang、C. Guan 和 H. Zhang，“BCI 竞赛 IV 数据集 2a 和 2b 上的滤波器组常见空间模式算法”，Frontiers Neurosci.，第 6 卷，第 1 期，第 39 页，2012 年。

[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.

[9] X. Chen, Y. Wang, S. Gao, T.-P. Jung 和 X. Gao, “滤波器组典型相关分析用于实现基于高速 SSVEP 的脑机接口”，J. Neural Eng.，第 12 卷，第 4 期，2015 年 8 月，文章编号 046008。

[10] C. Ieracitano、N. Mammone、A. Hussain 和 FC Morabito, “一种基于多模态机器学习的痴呆症脑电图记录自动分类新方法”，神经网络，第 123 卷，第 176-190 页，2020 年 3 月。

[11] A. Bhattacharyya、L. Singh 和 RB Pachori, “傅里叶-贝塞尔基于级数展开的经验小波变换用于非平稳信号分析》，数字信号处理，第 78 卷，第 185-196 页，2018 年 7 月。

[12] A. Bhattacharyya 和 RB Pachori, “多元方法用于使用经验小波变换进行患者特定 EEG 癫痫发作检测”，IEEE Trans. Biomed. Eng.，第 64 卷，第 9 期，第 2003-2015 页，2017 年 9 月。

[13] X. Wu、W.-L. Zheng、Z. Li 和 B.-L. Lu, “研究基于脑电图的功能连接模式以实现多模态情绪识别”，J. Neural Eng.，第 19 卷，第 1 期，2022 年 2 月，文章编号 016012。

[14] H. Göksu, “利用对数能量熵进行 BCI 导向脑电图分析小波包”，《生物医学信号处理控制》，第 44 卷，第 101-109 页，2018 年 7 月。

[15] K. He、X. Zhang、S. Ren 和 J. Sun, “深度残差学习用于图像识别”，载于 Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)，2016 年 6 月，第 770–778 页。

[16] RT Schirrmeister 等人, “基于卷积神经网络的深度学习用于脑电图解码和可视化的网络：脑电图分析中的卷积神经网络”，Hum. Brain Mapping，第 38 卷，第 11 期，第 5391-5420 页，2017 年 11 月。

[17] VJ Lawhern、AJ Solon、NR Waytowich、SM Gordon、CP Hung、BJ Lance, “EEGNet：基于 EEG 的脑机接口紧凑卷积神经网络”，J. Neural Eng.，第 15 卷，第 5 期，2018 年 10 月，文章编号 056013。

[18] S. Tortora、S. Ghidoni、C. Chisari、S. Micera 和 F. Artoni, “基于深度学习的 BCI，利用 LSTM 循环神经网络从 EEG 中解码步态”，J. Neural Eng.，第 17 卷，第 4 期，2020 年 7 月，货号 046011。

[19] A. Shoeibi 等人, “通过脑电图信号自动诊断精神分裂症使用 CNN-LSTM 模型”，Frontiers Neuroinform.，第 15 卷，2021 年 11 月，文章编号 777977。

[20] A. Vaswani 等人, “你只需要注意力”，载于 Proc. Adv. Neural Inf. Process. Syst.，第 30 卷。美国纽约州红钩：Curran Associates，2017 年，第 1-11 页。

[21] J. Xie 等人, “一种基于 Transformer 的方法，结合深度学习网络和时空信息进行原始脑电图分类”，IEEE Trans. Neural Syst. Rehabil. Eng.，第 30 卷，第 2126-2136 页，2022 年。

[22] Y. Song、X. Jia、L. Yang 和 L. Xie, “基于 Transformer 的时空特征学习用于 EEG 解码”，2021 年 6 月。[在线]。可访问网址：https://arxiv.org/abs/2106.11170 [23] S. Bagchi 和 DR Bathula, “EEG-ConvTransformer 用于单次试验基于脑电图的视觉刺激分类”，《模式识别》，第 129 卷，2022 年 9 月，文章编号 108757。

[24] Y. Zheng、X. Zhao 和 L. Yao, “基于 Copula 的 EEG 变换器评估立体 3D 引起的视觉不适”，《生物医学信号过程控制》，第 77 卷，2022 年 8 月，文章编号 103803。

[25] F. Lotte 等人, “基于脑电图的分类算法综述“脑机接口：10 年更新”，《神经工程学杂志》，第 15 卷，第 3 期，2018 年 6 月，文章编号 031005。

[26] PV 和 A. Bhattacharyya, “基于时间的人类情绪识别——多变量脑电信号的频率分析”， Knowl.-Based Syst.，第 238 卷，2022 年 2 月，文章编号 107867。

[27] A. Bhattacharyya、RK Tripathy、L. Garg 和 RB Pachori, “一种用于计算脑电图频谱和时间复杂度以实现人类情绪识别的新型多变量多尺度方法”，IEEE Sensors J.，第 21 卷，第 3 期，第 3579-3591 页，2021 年 2 月。

[28] S. Sakhavi、C. Guan 和 S. Yan, “学习时间信息使用卷积神经网络的脑机接口”，IEEE Trans. Neural Netw. Learn. Syst.，第 29 卷，第 11 期，第 5619-5629 页，2018 年 11 月。

[29] X. Shan、J. Cao、S. Huo、L. Chen、PG Sarrigiannis 和 Y. Zhao, “基于脑电图脑功能连接成像的时空图卷积网络用于阿尔茨海默病分类”，《人类。脑图谱》，第 43 卷，第 17 期，第 5194-5209 页，2022 年 6 月。

[30] X. Hong 等, “运动机器人动态关节域自适应网络图像分类”，IEEE 神经系统康复工程学报，第 29 卷，第 556-565 页，2021 年。

[31] W. Huang、W. Chang、G. Yan、Z. Yang、H. Luo 和 H. Pei, “基于脑电图的使用具有局部重参数化技巧的卷积神经网络进行运动想象分类”，专家系统应用，第 187 卷，2022 年 1 月，文章编号 115968。

[32] AM Roy, “一种高效的多尺度 CNN 模型，具有内在脑机接口中运动想象脑电图对象分类的特征整合”，《生物医学信号处理控制》，第 74 卷，2022 年 4 月，货号 103496。

[33] A. Dosovitskiy 等人, “一张图片胜过 16×16 个单词：Transformers 用于大规模图像识别”，载于 Proc. Int. Conf. Learn. Represent.，2022 年 3 月，第 1-21 页。

[34] D. Kostas、S. Aroca-Ouellette 和 F. Rudzicz, “BENDR：使用 Transformer 和对比自监督学习任务，从大量 EEG 数据中学习”，《Frontiers Hum. Neurosci.》，第 15 卷，第 253 页，2021 年 6 月。

[35] J. Liu, L. Zhang, H. Wu, 和 H. Zhao, “脑电图情绪转换器”认可”，2021 年 10 月。[在线]。可访问网址：https://arxiv.org/abs/2110.06553

[36] W.-L. Zheng 和 B.-L. Lu, “研究关键频带and channel for EEG-based emotions identification with deep neural networks（基于脑电图的情绪识别与深度神经网络通道），” IEEE Trans. Auton. Mental Develop.，第 7 卷，第 3 期，第 162-175 页，2015 年 9 月。

[37] F. Lotte, “最小化或抑制基于振荡活动的脑机接口中校准时间的信号处理方法”，Proc. IEEE，第 103 卷，第 6 期，第 871-890 页，2015 年 6 月。

[38] R. Mane 等人, “FBCNet：一种多视图卷积神经网络脑机接口网络”，2021 年 3 月。[在线]。链接：https://arxiv.org/abs/2104.01233

[39] H. Zhao, Q. Zheng, K. Ma, H. Li, 和 Y. Zheng, “深度表征-基于域自适应的非平稳脑电图分类”，IEEE 神经网络学习系统汇刊，第 32 卷，第 2 期，第 535-545 页，2021 年 2 月。

[40] W.-L. Zheng、J.-Y. Zhu 和 B.-L. Lu, “识别通过脑电图识别情绪的时间”，IEEE 情感计算汇刊，第 10 卷，第 3 期，第 417-429 页，2019 年 7 月。

[41] T. Song, W. Zheng, P. Song 和 Z. Cui, “使用动态图卷积神经网络进行脑电图情绪识别”，IEEE Trans. Affect。Comput.，第 11 卷，第 3 期，第 532-541 页，2020 年 7 月。

[42] Y. Li, W. Cheng, L. Wang, Y. Zong, and Z. Cui, “从区域到全球brain：一种用于脑电图情绪识别的新型分层时空神经网络模型》，IEEE 情感计算汇刊，第 13 卷，第 2 期，第 568-578 页，2022 年 4 月。

[43] P. Zhong、D. Wang 和 C. Miao, “基于脑电图的情绪识别使用正则化图神经网络”，IEEE 情感计算汇刊，第 13 卷，第 3 期，第 1290-1301 页，2022 年 7 月。

[44] L. van der Maaten 和 G. Hinton, “使用 t-SNE 可视化数据”J. Mach. Learn. Res.，第 9 卷，第 2579-2605 页，2008 年 11 月。

[45] RR Selvaraju、M. Cogswell、A. Das、R. Vedantam、D. Parikh 和 D. Batra, “Grad-CAM：通过基于梯度的定位从深度网络实现视觉解释”，载于 Proc. IEEE Int. Conf. Comput. Vis.（ICCV），2017 年 10 月，第 618-626 页。

[46] A. Schnitzler、S. Salenius、R. Salmelin、V. Jousmäki 和 R. Hari, “初级运动皮层在运动意象中的参与：一项神经磁学研究”，《神经影像》，第 6 卷，第 3 期，第 201-208 页，1997 年 10 月。