ADFCNN: Attention-Based Dual-Scale Fusion Convolutional Neural Network for Motor Imagery Brain–Computer Interface

Wei Tao, Ze Wang, Chi Man Wong, Ziyu Jia, Chang Li, *Member, IEEE*, Xun Chen, *Senior Member, IEEE*, C. L. Philip Chen, *Fellow, IEEE*, and Feng Wan, *Senior Member, IEEE*

***Abstract*— Convolutional neural networks (CNNs) have been successfully applied to motor imagery (MI)-based brain–computer interface (BCI). Nevertheless, single-scale CNN fail to extract abundant information over a wide spectrum from EEG signals, while typical multi-scale CNNs cannot effectively fuse information from different scales with concatenation-based methods. To overcome these challenges, we propose a new scheme equipped with attention-based dual-scale fusion convolutional neural network (ADFCNN), which jointly extracts and fuses EEG spectral and spatial information at different scales. This scheme also provides novel insight through self-attention for effective information fusion from different scales. Specifically, temporal convolutions with two different kernel sizes identify EEG μ and β rhythms, while spatial convolutions at two different scales generate global and detailed spatial information, respectively, and the self-attention mechanism performs feature fusion based on the internal similarity of the concatenated features extracted by the dual-scale CNN. The proposed scheme achieves the superior performance compared with state-of-the-art methods in subject-specific motor imagery recognition on BCI Competition IV dataset 2a, 2b and OpenBMI dataset, with the cross-session average classification accuracies of 79.39% and significant improvements of 9.14% on BCI-IV2a, 87.81% and 7.66% on BCI-IV2b, 65.26% and 7.2% on OpenBMI dataset, and the within-session average classification accuracies of 86.87% and significant improvements of 10.89% on BCI-IV2a, 87.26% and 8.07% on BCI-IV2b, 84.29% and 5.17% on OpenBMI dataset, respectively. What is more, ablation experiments are conducted to investigate the mechanism and demonstrate the effectiveness of the dual-scale joint temporal-spatial CNN and self-attention modules. Visualization is also used to reveal the learning process and feature distribution of the model.**

***Index Terms*— Convolutional neural networks (CNNs), motor imagery (MI), brain–computer interface (BCI), self-attention mechanism.**

Manuscript received 17 May 2023; revised 15 August 2023, 2 October 2023, and 14 November 2023; accepted 8 December 2023. Date of publication 13 December 2023; date of current version 16 January 2024. This work was funded in part by The Science and Technology Development Fund, Macau SAR (File no. 0045/2019/AFJ, 0022/2021/APD, 0007/2023/RIC, 0024/2023/ITP1, 0024/2023/R1A1); in part by University of Macau (File no. MYRG2017-00207-FST, MYRG-2022-00197-FST); and in part by the Guangdong Basic and Applied Basic Research Foundation (File no. 2023A1515010844). *(Corresponding author: Feng Wan.)*

Wei Tao, Chi Man Wong, and Feng Wan are with the Department of Electrical and Computer Engineering, the Faculty of Science and Technology, the Centre for Cognitive and Brain Sciences, the Centre for Artificial Intelligence and Robotics, and the Institute of Collaborative Innovation, University of Macau, Taipa, Macau (e-mail: yc07466@umac.mo; chiman465@gmail.com; fwan@um.edu.mo).

Ze Wang is with the Macao Centre for Mathematical Sciences, and the Respiratory Disease AI Laboratory on Epidemic Intelligence and Medical Big Data Instrument Applications, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau (e-mail: zwang@must.edu.mo).

Ziyu Jia is with the Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jia.ziyu@outlook.com).

Chang Li is with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: changli@hfut.edu.cn).

Xun Chen is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: xunchen@ustc.edu.cn).

C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: Philip.Chen@ieee.org).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2023.3342331, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2023.3342331

I. Introduction

BRAIN–COMPUTER interface (BCI) have emerged as a promising augmentative communication and control technology [1], [2], [3], [4], [5]. Over the years, several paradigms of electroencephalogram (EEG)-based BCI have been developed, including steady-state visual evoked potentials (SSVEP) [6], event related potentials (ERP) [7], emotion [8], and motor imagery (MI) [9], [10], [11], [12], [13]. Among these paradigms, MI-based BCI has garnered significant attention, as it enables decoding of users' motor intentions from EEG signals. It has been successfully applied in various fields, such as stroke rehabilitation [14], wheelchair control [15], cursor control [16], among others.

Despite the advancements in the field of BCI, accurately decoding motor intentions from EEG signals remains a challenge due to the complex characteristics of EEG, such as low signal-to-noise ratio (SNR), non-stationarity, low spatial resolution, high temporal resolution, and inter-individual variability [10]. Currently, two main methods are widely employed for MI-based BCI decoding: traditional machine learning and deep learning. Traditional machine learning methods generally involve two distinct steps of feature extraction and feature classification [17]. These techniques include signal processing

techniques such as fast Fourier transform (FFT) [18], common spatial pattern (CSP) [19], wavelet transform (WT) [20], and short-time Fourier transform (STFT) [21], which can extract frequency, spatial-frequency or time-frequency features from EEG signals. Supervised classification algorithms such as random forest (RF) [22], support vector machine (SVM) [23], and linear discriminant analysis (LDA) [24], and unsupervised learning techniques such as K-nearest neighbor analysis (KNN) [25], are applied to classify these features. However, traditional machine learning methods are labor-intensive, requiring significant expert knowledge, thus limiting classification performances. Conversely, deep learning (DL) methods have recently achieved immense successes in the field of BCI owing to their powerful representation learning capabilities [26]. Moreover, DL-based methods can be implemented with an end-to-end framework, combining feature extraction and classification into a single, integrated, and convenient scheme.

Currently, convolutional neural networks (CNNs) have emerged as important deep learning structures in MI-based BCI, owing to their powerful representation learning capabilities. Several studies have investigated the impacts of different CNN parameters, such as the convolution manner, kernel size, number of kernels, and layers. For instance, Schirrmeister et al. proposed two different CNN-based architectures, namely Shallow ConvNet and Deep ConvNet, to classify MI end-to-end and found that the depth of the CNN significantly influenced its performance [27]. Lawhern et al. applied the separable convolution operation in CNN, to develop EEGNet, a general BCI classification framework that successfully classified various tasks [28]. Hermosilla et al. experimented with a shallow CNN that implemented varied numbers and sizes of kernels to improve MI classification performance [29]. In addition, some studies have also leveraged the unique characteristics of EEG signals when designing CNN for the MI classification task. For example, Mane et al. proposed the FBCNet based on filter-bank CSP, wherein the CNN extracted information from multi-frequency-band signals to classify MI tasks [30]. Wang et al. proposed a novel, lightweight interactive frequency convolutional neural network named IFNet, which can further explore the cross-frequency interactions for enhancing the representation of MI characteristics [31]. Lee and Choi proposed a novel pipeline, first extracting time-frequency representations by continuous wavelet transform (CWT) and then applying a CNN on the CWT representations for MI classification task [32].

Recent advances in single-scale CNNs have shown promising results in MI-based BCI. However, these models have limited ability in effectively capturing the abundant information present in EEG data. As a result, recent studies are exploring the use of multi-scale CNNs, which employ multiple convolutional scales to better extract information from EEG signals. For instance, Dai et al. proposed a hybrid-scale CNN architecture to consider the differences of convolution scales on individuals [33]. Ko et al. proposed a novel deep multi-scale neural network to extract feature representations from multiple frequency/time ranges and to discover relationships among electrodes [34]. Zhao et al. proposed a multi-branch 3D CNN, mainly consisting of three scale convolutions to extract the temporal-spatio information from raw signals [35]. Nevertheless, single-scale CNNs fail to extract abundant information across a wide spectrum from EEG signals, while conventional multi-scale CNNs overlook the fusion of different scale information.

To address these limitations, this paper proposes a new scheme equipped with attention-based dual-scale fusion convolutional neural network, which can jointly extract and fuse EEG spectral and spatial information at different scales, leading to superior performance compared with state-of-the-art methods. The main contributions are as follows:

- To obtain different-scale spectral and spatial information by different types of temporal and spatial convolutions, we propose a novel dual-scale temporal-spatial CNN to jointly identify EEG μ and β rhythms, as well capture global and detailed spatial information.
- To explore the implicit information of fused features in multi-scale CNNs, we consider the internal similarity of dual-scale features extracted from the dual-scale joint temporal-spatial CNN, and apply a self-attention mechanism to adaptively enhance the flexibility of the fusion feature.
- To verify the effectiveness of our proposed method, we conduct comprehensive experiments with the proposed scheme on three public datasets for subject-specific MI-based BCI and achieve promising results. Specifically, the cross-session average classification accuracies are 79.39% on BCI-IV2a, 87.81% on BCI-IV2b, 65.26% on OpenBMI dataset, while the within-session average classification accuracies are 86.87% on BCI-IV2a, 87.26% on BCI-IV2b, and 84.29% on OpenBMI dataset, respectively.

The code is publicly available at https://github.com/UM-Tao/ADFCNN-MI.

II. Related Work

In the past decade, CNNs have demonstrated remarkable successes in the field of computer vision, owing to their ability to capture both global and local features from image signals through convolution operations [36]. When processing EEG signals, convolution in the time or spatial domain of EEG signals involves weighting the EEG signals by the sliding kernel. Therefore, CNNs have also been utilized in the field of BCI to extract relevant spectral and spatial information from EEG signals via convolution operations. Generally, temporal convolutions operate one-dimension (1)-D) convolution along the temporal dimension to extract EEG spectral information. A large kernel size allows for capturing low-frequency information, while a small kernel size allows for capturing high-frequency information [37]. Spatial convolutions operate the 1-D convolution along the channel dimension to extract spatial information from EEG signals, wherein kernels can be considered as spatial filters, and the learned weights are indicative of the information from different electrodes. In addition, there are two types of convolution operations that are commonly used: standard convolution operates on all feature

maps, and enables the sharing of weights among feature maps, while separable convolution operates on each feature map one by one, which provides sparse information with fewer parameters. Moreover, CNNs can be categorized as single-scale and multi-scale depending on the application of different-scale convolutions. Single-scale CNNs extract spectral and spatial information from EEG signals using single-scale temporal and spatial convolution, but the efficacy of single-scale convolution may vary from subject to subject, session to session and even time point to time point in EEG classification tasks. Compared to single-scale CNNs, multi-scale CNNs leverage convolutions on multiple scales to gain more comprehensive insights from EEG signals. For instance, Dai et al. explored three types of temporal convolutions and developed a multi-scale CNN architecture to enhance MI classification accuracy [33]. Zhao et al. converted EEG signals into 3D representation and employed a 3-branch 3D CNN to address MI classification tasks [35]. Jia et al. proposed a multi-branch multi-scale CNN to extract different-scale spectral information from EEG signals, thereby improving MI classification performance [38].

Although multi-scale CNNs extract more information compared with single-scale CNNs, it is challenging to effectively fuse different-scale information in the feature fusion module for multi-scale CNNs. Conventional multi-scale CNNs often utilize a straightforward concatenation of all the features, disregarding the internal connections between and among different-scale features. However, recent advances in attention mechanism have allowed for the successful extraction of implicit information from features [39], [40], [41]. For instance, channel-wise attention extracts the importance information among channels through allocating weights to different channels [40]. Temporal attention is used to capture long-range temporal dependency of time series data [41]. Moreover, self-attention mechanism, which considers the similarity among the feature vectors (*i.e.*, query (Q), key (K), and value (V)), can be used to obtain the self-attention weights by softmax function calculation. What is more, self-attention mechanism has been widely used in the field of BCI due to its ability to explore the time dependency of EEG slices. For instance, Zhang et al. proposed a convolutional recurrent attention model, which combines self-attention mechanism with CNN and Long Short-term Memory (LSTM) to extract the intrinsic time-spatial dependency among different time slices [42]. Xie et al. proposed a transformer-based deep model with the self-attention mechanism to extract implicit spatial information from raw EEG signals [43]. Considering the ability of self-attention mechanisms to explore the internal correlations among features, it can be incorporated to improve the flexibility of fusion features with multi-scale CNNs, leading to an enhanced classification performance of the deep learning model.

III. Materials and Methods

A. Dataset Description

In this study, we evaluated the effectiveness of the proposed model using three widely-recognized public datasets. Each dataset differs in the number of subjects, electrodes, signal qualities and experimental setup. The details of each dataset are presented as follows:

1) BCI Competition IV 2a Dataset (BCI-IV2a) [44]: It was collected from nine healthy subjects with a sampling rate of 250 Hz. For each subject, 576 trials from two EEG sessions, recorded from 22 Ag/AgCl electrodes according to the international 10-20 system, were obtained. The dataset contained four types of MI tasks, including those for left-hand, right-hand, both feet, and tough. Each trial lasted four seconds during the MI period.

2) BCI Competition IV 2b Dataset (BCI-IV2b) [45]: It comprised EEG recordings obtained from nine subjects at a sampling rate of 250 Hz, using three electrodes located in positions C3, Cz, and C4 according to the international 10-20 system. Two MI tasks (left-hand vs. right-hand) were performed and each trial had an imagery duration of 4s. The first two sessions comprised 400 trials, and the remaining three sessions had 320 trials each. The first two sessions were processed without feedback, whereas the remaining three sessions were processed with feedback.

3) OpenBMI Dataset [46]: It was collected from fifty-four subjects using 62 Ag/AgCl electrodes at a sampling rate of 1000 Hz. The dataset consisted of a hybrid of three paradigms (MI, ERP, and SSVEP). It contained two MI tasks (left-hand vs. right-hand), and contained two sessions per subject with 200 trials per session, with each trial lasting 4s for the MI paradigm.

B. EEG Data Preprocessing

In preprocessing, each EEG trial can be first described $\mathbf{X} \in \mathbb{R}^{C\times T}$ where C is the number of EEG electrode nodes, T is the number of sampling points. For electrode, we consider all electrodes for two BCI competition datasets where C is set to 22 and 3, and we consider these electrodes related to motor region for OpenBMI dataset where C is set to 20 according to the [46]. For sampling rate, we down-sample raw data from 1000 Hz to 250 Hz for openBMI dataset to keep the same sampling rate of three datasets. For the period, we consider the 0s-3s after the cue in the imagery period for processing as suggested by [27]. Later, we apply a bandpass filter to obtain filtered EEG signals between 0-40 Hz following the recommendation in [27]. Moreover, the influences due to different choices of periods and frequency bands have been further discussed as shown in Tables S1-S4 of the supplementary file. In addition, we employ the electrode-wise exponential moving standardization [12], [47], [48] to obtain standardized EEG data $\mathbf{X}' \in \mathbb{R}^{C\times T}$ as shown in the following formulas:

$$m_t = \alpha \cdot \text{mean}(x_t) + (1-\alpha)\cdot m_{t-1}, \tag{1}$$

where $x_t \in \mathbb{R}^{C\times 1}$ denotes the value at time t of $\mathbf{X}$, m_t denotes the moving mean value at time t, the α denotes the decay factor and is set to 0.001.

$$v_t = \alpha \cdot (m_t - x_t)^2 + (1-\alpha)\cdot v_{t-1}, \tag{2}$$

where v_t denotes the moving variance value at time t,

$$x'_t = (x_t - m_t)/\sqrt{v_t}, \tag{3}$$

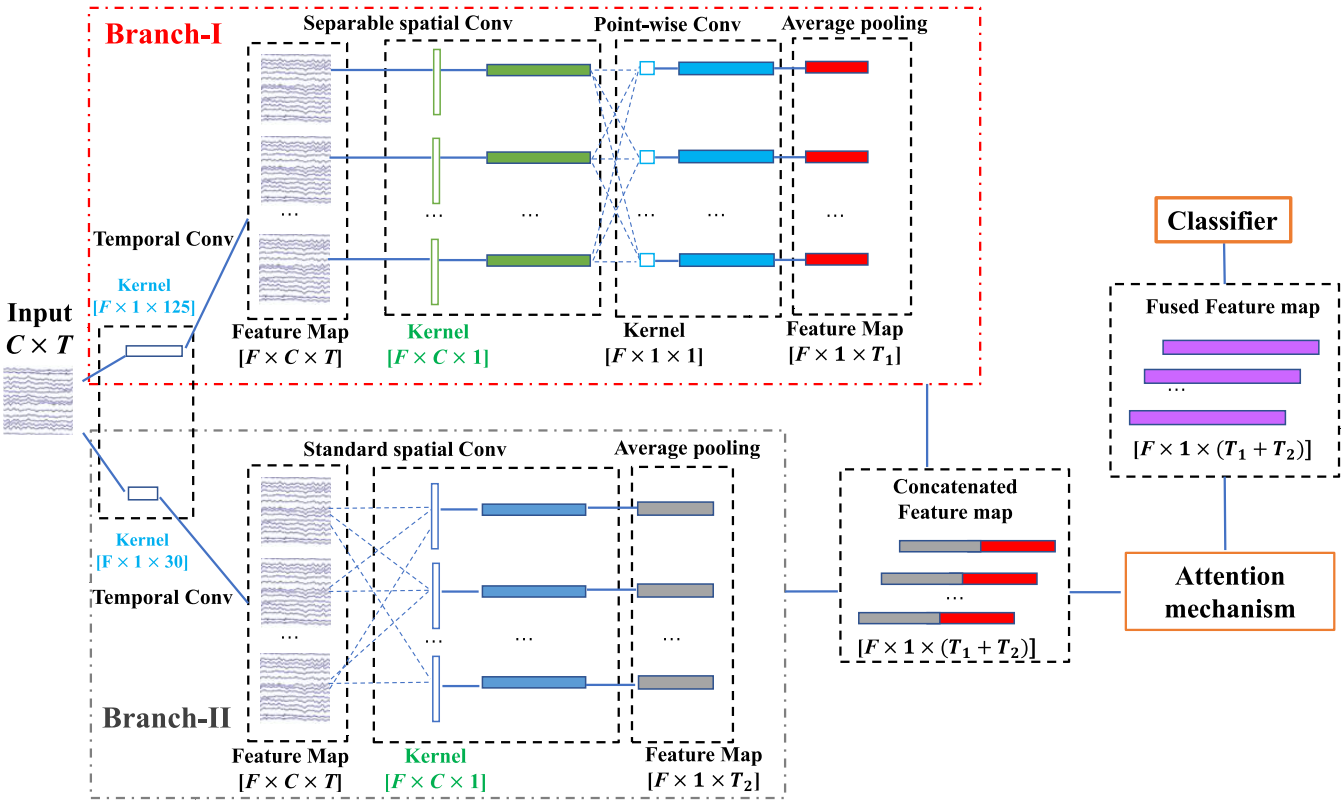

Fig. 1. Framework of the proposed attention-based dual-scale fusion convolutional neural network for MI-EEG classification.

TABLE I
THE PARAMETERS OF ADFCNN ARCHITECTURE

Module	Layer	Size	Params	Output Shape
Branch-I	Temporal convolution	$[F \times 1 \times 125]$	$F \times 125$	$[F \times C \times T]$
	Separable spatial convolution	$[F \times C \times 1]$	$F \times C$	$[F \times 1 \times T]$
	Point-wise convolution	$[F \times 1 \times 1]$	$F \times F$	$[F \times 1 \times T]$
	Average pooling	$[1 \times 32]$	/	$[F \times 1 \times T_1]$
Branch-II	Temporal convolution	$[F \times 1 \times 30]$	$F \times 30$	$[F \times C \times T]$
	Standard spatial convolution	$[F \times C \times 1]$	$C \times F \times F$	$[F \times 1 \times T]$
	Average pooling	$[1 \times 75]$	/	$[F \times 1 \times T_2]$
Fusion	Concatenation	/	/	$[F \times 1 \times (T_1 + T_2)]$
	Attention-based	/	$C \times C \times 3$	$[F \times 1 \times (T_1 + T_2)]$
Classifier	Dense	/	$N \times F \times (T_1 + T_2)$	$[N \times 1]$

C = number of channels, T = number of time points, F = number of convolution kernel,
T_1 = dimension of features learned by Branch I, T_2 = dimension of features learned by Branch II, N = number of classes

where x'_t denotes the value at time t of standardized EEG data $\mathbf{X}'$.

C. Model Architecture

Temporal convolutions with large kernel sizes are able to extract broader frequency information from EEG signals, while those with small kernel sizes can extract higher frequency information. Similarly, spatial convolutions with separable manner can extract global and detailed spatial information, while standard manner can extract abundant and detailed spatial information. To take advantage of the different-scale temporal and spatial convolutions, we introduce a dual-scale joint temporal-spatial CNN module capable of capturing multi-scale spectral and spatial information of EEG signals. Additionally, we design an attention-based feature fusion module to effectively fuse the features extracted by the dual-scale CNN, thereby exploring the similarity of concatenated features and improving the flexibility of the fusion feature. The proposed scheme is illustrated in Fig. 1, where the input is 2D EEG signals. A dual-branch temporal-spatial CNN is used to extract both spectral and spatial features from the EEG signal. Subsequently, a self-attention mechanism is employed to fuse the concatenated features. Finally, a dense layer is used as the classifier to calculate the classification result. The detailed structure of this model is presented in Table I, which includes information such as modules, layers, kernel size, number of trainable parameters and output shape.

In the first temporal-spatial CNN branch (Branch-I), which focuses on capturing larger-scale spectral and spatial information, we incorporated both large-scale temporal convolution and separated spatial convolutions. Temporal convolution layer employs F large-scale kernels with size $[1 \times 125]$ to extract filtered feature maps from raw EEG signals as suggested by [28]. These temporal kernels with large size mainly capture

low-frequency information. Subsequently, separable spatial convolution layer employs F spatial kernels with size $[C \times 1]$ to extract global spatial information of the feature map one by one. This operation not only helps to reduce the number of trainable parameters in convolution operations, but also adds a layer of sparsity to the feature map. To combine these separated spectral-spatial features, a point-wise convolution layer with kernel size $[1 \times 1]$ is then used to integrate the separated spatial features and synthesize them into a comprehensive representation. Finally, an average pooling layer with size $[1 \times 32]$ reduces dimension of features following the recommendation in [28]. The learning process of Branch-I can be summarized as follows:

$$\mathbf{X}_{T_1}^r = \text{TConv}_r\left(\mathbf{X}'\right), r \in [1, 2, \ldots, F], \tag{4}$$

$$\mathbf{X}_{S_1}^r = \text{Separable\_SConv}_r\left(\mathbf{X}_{T_1}^r\right), r \in [1, 2, \ldots, F], \tag{5}$$

$$\mathbf{X}_P = \text{Pointwise\_Conv}\left(\mathbf{X}_{S_1}\right), \tag{6}$$

$$\mathbf{X}_F^1 = \text{AveragePool}\left(\mathbf{X}_P\right), \tag{7}$$

where $\mathbf{X}'$ is the input, TConv_r is the temporal convolution with r-th temporal kernel, $X_{T_1}^r$ is the output of large-scale temporal convolution, Separable_SConv_r is the separable spatial convolution with r-th spatial kernel, $\mathbf{X}_{S_1}^r$ is the output of separable spatial convolution, Ponitwise_Conv is the point-wise convolution to mix the feature maps by separable manner and X_P is the output of point-wise convolution. AveragePool is average pooling operation. $\mathbf{X}_F^1$ denotes the learned feature map in Branch-I with size $[F \times 1 \times T_1]$.

In the second temporal-spatial CNN branch (Branch-II), which aims to capture smaller-scale spectral and standard spatial information. F small-scale temporal convolution kernels with size $[1 \times 30]$ are employed to mainly extract feature maps at high-frequency information as suggested by [27]. A standard spatial convolution layer learns global spatial information from all feature maps, utilizing $F \times F$ standard convolution kernels with size $[C \times 1]$. An average pooling layer with size $[1 \times 75]$ is applied to reduce the feature dimension and preserve large-scale temporal information of learned features according to [27]. The learning process of Branch-II can be formalized as follows:

$$\mathbf{X}_{T_2}^r = \text{TConv}_r\left(\mathbf{X}'\right), r \in [1, 2, \ldots, F], \tag{8}$$

$$\mathbf{X}_{S_2} = \text{Standard\_SConv}\left(\mathbf{X}_{T_2}\right), \tag{9}$$

$$\mathbf{X}_F^2 = \text{AveragePool}\left(\mathbf{X}_{S_2}\right), \tag{10}$$

where $\mathbf{X}'$ denotes the same input as in Branch-I, $\mathbf{X}_{T_2}^r$ is the output of small-scale temporal convolution with r-th temporal kernel, standard_SConv is the standard spatial convolution and $\mathbf{X}_{S_2}$ is the output of standard spatial convolution, $\mathbf{X}_F^2$ denotes the learned feature map in Branch-II with size $[F \times 1 \times T_2]$.

In the feature fusion module, we first concatenate features extracted by dual-scale joint temporal-spatial CNN module as follows:

$$\mathbf{X}_F = \text{Concat}\left(\mathbf{X}_F^1, \ \mathbf{X}_F^2\right), \tag{11}$$

where Concat denotes the concatenation operation and $\mathbf{X}_F$ is concatenated features with size $[F \times 1 \times (T_1 + T_2)]$. To further explore the implicit information of concatenated feature map $\mathbf{X}_F$, we apply the self-attention to calculate the similarity within $\mathbf{X}_F$ and reallocate attention weights to F channels of $\mathbf{X}_F$ [39]. The self-attention mechanism is shown in Fig. 2, where $H = 1$ and $W = (T_1 + T_2)$ in concatenated feature map X_F. To obtain the internal similarity of features, the $\mathbf{X}_F$ is multiplied into three different weight matrices by linear transformation to obtain the queries (Q), keys (K), and values (V) as follows:

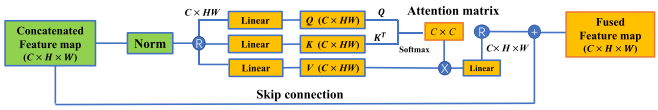

Fig. 2. The detailed architecture of self-attention mechanism.

$$Q, K, V = \text{Linear}\left(\mathbf{X}_F, \ W^Q, \ W^K, \ W^V\right), \tag{12}$$

where Linear denotes the linear transformation with weights, W^Q, W^K, and W^V are trainable parameters in linear transformation. To obtain the attention score of concatenated features, the similarity of different features is first computed with dot products of the queries (Q) with keys (K). Then, the results of dot product are divided each by $\sqrt{d_k}$ and a softmax function is applied to obtain the attention matrix. Lastly, the attention matrix is considered as weights to allocate on values (V), the calculation can be described as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V, \tag{13}$$

where d_k is the dimension of Keys (K), it is set to 64 in our work. In addition, we also apply the skip-connection to keep the intrinsic information of concatenated features in feature fusion module.

Finally, we apply a dense layer with a softmax function to the output, yielding an M-dimensional feature vector. We employ cross-entropy as the loss function for model training, which is expressed as:

$$\mathcal{L} = -\frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{c=1}^{M} y \log(\hat{y}), \tag{14}$$

here M denotes the number of MI tasks, y and $\hat{y}$ denote the actual and predicted label, respectively. N_b represents the number of samples in one batch.

D. Simulation Setup

The proposed method is first compared with five representative deep learning baselines by experiments, including EEGNet [28], Deep ConvNet [27], Shallow ConvNet [27], FBCNet [30], and IFNet [31]. Then, we compare the proposed method with several state-of-the-art (SOTA) deep learning methods by literature including MCNN [47], MI-EEGNET [49], MSHCNN [50], MMCNN [38], Tensor-CSPNet [51], SHNN [52], and Conformer [53]. In the experiments for comparison, the hyperparameters include the training epoch, the batch size, the learning rate, and the weight decay. Due to that the learning rate and weight decay are significant for different deep models in network optimization,

we also conducted experiments to determine the optimal values of these two hyperparameters for each model using the BCI-IV2a dataset as shown in Figures S1 and S2 of the supplementary file. Consequently, we selected the optimal value of learning rate, e.g., 0.001 for each deep model, and selected different respective optimal value of weight decay for different methods, e.g., 0.075 for EEGNet, Shallow ConvNet, Deep ConvNet, and ADFCNN, 0.001 for FBCNet and IFNet. The rest hyperparameters keep consistent with different methods, such as training epoch, and the batch size are set to 1000, 16, respectively. In addition, all methods adopt Adam as an optimizer and cross-entropy loss as loss function. For platform, these methods are implemented in Python with Pytorch, and are trained on NVIDIA Tesla v100 GPU.

As an active BCI paradigm, motor imagery task requires subjects to actively engage in self-practice [54], the proposed method and compared methods are mainly evaluated in subject-specific MI-based BCI which refers to training the decoding model using the target subject's data. Subject-specific MI has the potential to have better decoding accuracy as the deep model can capture the individual's unique neural patterns [50], [51], [53]. Two common cases of subject-specific MI-based BCI are considered including within-session and cross-session scenarios. In the former, training and testing sets are taken from the same session, while in the latter, they are taken from distinct, independent sessions. For within-session case, we performed the five-fold cross-validation specifically on the data from the first session, and then reported the average accuracy across these five folds as our within-session result. For cross-session evaluation, we applied the same five-fold cross-validation procedure using the data from the first session, the reported results reflect the average performance of five models tested on the second session's data.

To statistically compare the classification results of the proposed method with those of five deep learning baselines in experiments, we employ Wilcoxon signed rank test [55] on all subjects between the proposed method and baselines. We then estimate the $p-$value to be less than 0.05, which indicates a statistically significant difference between them.

IV. Experimental Results

In this section, we first present detailed comparative experiments to demonstrate the influence of the critical parameter. Subsequently, we conduct experiments to validate the proposed model on three public datasets. In addition to compare our method to several popular deep learning techniques based on the classification performance, we also demonstrate the effectiveness of the critical modules of the proposed model through ablation experiments. Finally, we explore various visualization methods to gain insights into model interpretability.

A. Effect of Kernel Number

Conventional multi-scale CNNs typically construct the network with multiple branches and layers to extract more information from EEG signals [38], [47], [49]. However, too many trainable parameters can lead to overfitting of limited training data and therefore, the number of trainable parameters is an important factor to consider when designing deep models for BCIs. To this end, Table I presents the critical network parameters, namely the number of used electrodes and temporal convolution kernels, which influence the number of trainable parameters. Therefore, to search for the optimal number of temporal convolution kernels, we conducted an experiment on the BCI-IV2a dataset, which has the most used electrodes compared to the other two datasets. The results of the effect of the kernel number, total parameters, and classification performance are illustrated in Fig. 3. Based on the experimental results, we set the number of temporal convolution kernels F to eight in the following simulations.

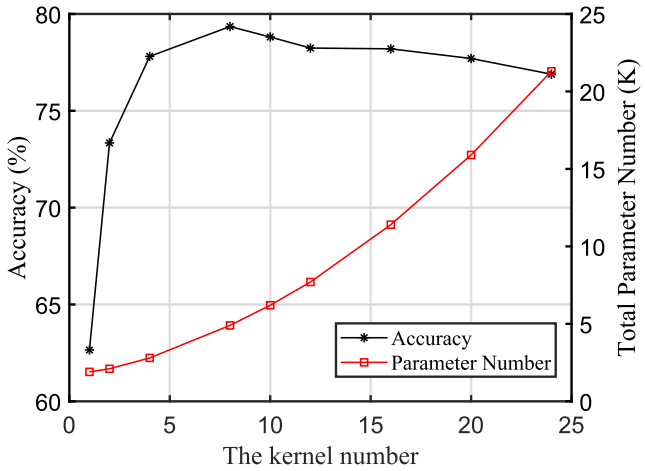

Fig. 3. Effect of kernel number on the total parameter number and classification performance in the proposed method.

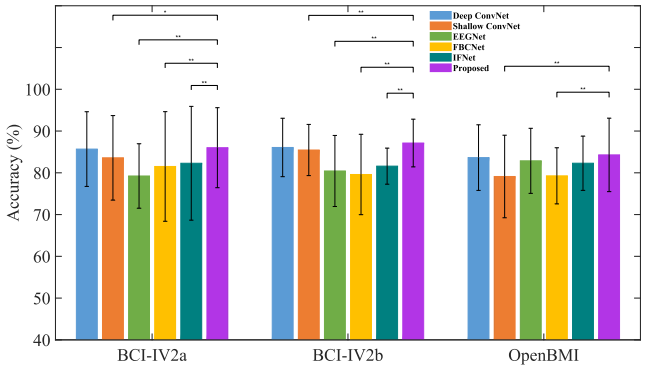

Fig. 4. Within-session classification performance comparison of the proposed method and other methods on three datasets, * : $p < 0.05$, ** : $p < 0.01$.

B. Classification Performance

Fig. 4 illustrates the average classification results of the proposed method and compared methods in the within-session case of subject-specific MI-based BCI. It is observed from Fig. 4 that ADFCNN achieves the highest classification accuracy on three datasets. The proposed method has significant improvements compared with four methods on two BCI competition datasets ($p < 0.05$) and performs significantly better than two baselines on OpenBMI dataset ($p < 0.05$). Fig. 5 presents the average classification results of the proposed method as well as compared methods in the cross-session case of subject-specific MI-based BCI. It can be found that the classification results of most methods have

TABLE II
THE VARIANT MODELS OF ABLATION EXPERIMENT

Component / Model	Temporal convolution layer	Separable spatial convolution layer	Standard spatial convolution layer	Average pooling layer	Concatenation layer	Self-attention layer
ADFCNN-Branch I	✓	✓	×	✓	×	×
ADFCNN-Branch II	✓	×	✓	✓	×	×
DFCNN	✓	✓	✓	✓	✓	×
ADFCNN	✓	✓	✓	✓	✓	✓

TABLE III
COMPARISON OF THE PROPOSED METHOD WITH FOUR SOTA METHODS IN CROSS-SESSION CASE ON BCI COMPETITION DATASETS

BCI-IV Dataset	Classification Accuracy(%)				
	Multi-scale CNN with concatenation fusion				Multi-scale CNN with attention fusion
	MCNN [47] 2a	MI-EEGNET [49] 2a	MSHCNN [50] 2b	MMCNN [38] 2b	Proposed 2a / 2b
Sub 1	90.21	80.91	86.80	84.90	87.15 / 79.37
Sub 2	63.40	60.76	77.94	70.40	61.45 / 72.50
Sub 3	89.35	87.46	65.97	75.50	93.75 / 82.81
Sub 4	71.16	57.23	97.97	96.30	75.69 / 96.25
Sub 5	62.82	77.39	93.24	92.40	75.34 / 99.37
Sub 6	47.66	66.32	88.88	86.30	65.27 / 84.68
Sub 7	90.86	88.92	86.80	87.60	88.54 / 93.43
Sub 8	83.72	84.34	82.89	84.20	82.29 / 95.31
Sub 9	82.32	94.09	86.80	81.80	85.06 / 86.56
Average	75.72	77.49	85.25	84.40	**79.39 / 87.81**

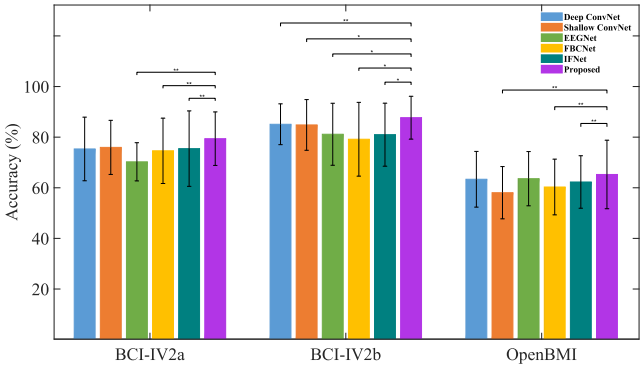

Fig. 5. Cross-session classification performance comparison of the proposed method and other methods on three datasets, * : $p < 0.05$, ** : $p < 0.01$.

decreased compared with those in the within-session case, due to differences caused by equipment and subject in the cross-session case. Nevertheless, the proposed method still has better classification results on three datasets and improves the results significantly compared with baselines ($p < 0.05$). These experiment results demonstrate that the proposed method has a superior classification performance for subject-specific MI-based BCI in both within-session and cross-session cases.

After careful comparison with five representative deep learning baselines by experiments, we also compare our proposed method with the reported results of four deep learning methods employing multi-scale CNNs with concatenation fusion, as shown in Table III. These multi-scale CNN methods include MCNN [47], MI-EEGNET [49], MSHCNN [50] and MMCNN [38]. MCNN is a multi-layer CNN method that fuses CNNs with different characteristics and architectures through feature concatenation, while MI-EEGNET is a Convnet based on the concatenation of Inception and Xception architectures. MSHCNN is a multi-scale hybrid CNN that concatenates one-dimensional and two-dimensional convolutions, and MMCNN is a multi-branch multi-scale network with concatenation. From Table III, it is evident that our proposed method achieves superior classification performance compared to these existing multi-scale CNN methods with concatenation fusion. In addition, we compared our proposed method with three latest deep learning methods including Tensor-CSPNet [51], SHNN [52], and Conformer [53]. Tensor-CSPNet is a novel geometric deep learning framework that exploits the temporo-spatio-frequency features of EEG signals. SHNN is a SincNet-based hybrid neural network that automatically filters data and extracts spatial, spectral, and temporal features from EEG. Conformer is a compact convolutional transformer network that combines local features and global features of EEG signals. Finally, We summarize the reported results of the five baselines, four multi-scale CNN methods, three latest methods, and our proposed method on two BCI competition datasets, as presented in Table IV. Overall, our proposed ADFCNN method outperforms existing multi-scale CNN methods with concatenation fusion and shows promising results compared to recent deep learning approaches for EEG classification.

C. Ablation Experiments

To study the effect of different components in the proposed model, we perform ablation experiments on three datasets and design the following variant models as shown in Table II:

TABLE IV
PERFORMANCE COMPARISON WITH RELATED WORKS

Related works	Methods	Within-session classification accuracy (%)		Cross-session classification accuracy (%)	
		BCI-IV2a	BCI-IV2b	BCI-IV2a	BCI-IV2b
Schirrmeister *et al.* 2017 [27]	Deep ConvNet	72.20	83.49	72.22	80.00
	Shallow ConvNet	76.47	85.51	73.70	84.83
Lawhern *et al.* 2018 [28]	EEGNet	74.15	83.96	73.15	80.15
Amin *et al.* 2019 [47]	MCNN	-	-	75.72	-
Mane *et al.* 2020 [30]	FBCNet	79.03	-	76.20	
Riyad *et al.* 2021 [49]	MI-EEGNET	-	-	77.49	-
Jia *et al.* 2021 [38]	MMCNN	-	-	-	84.40
Ju *et al.* 2022 [51]	Tensor-CSPNet	75.98	-	72.96	-
Liu *et al.* 2022 [52]	SHNN	-	-	74.26	83.49
Song *et al.* 2022 [53]	Conformer	-	-	78.66	84.63
Tang *et al.* 2023 [50]	MSHCNN	-	-	-	85.25
Wang *et al.* 2023 [31]	IFNet	-	-	78.21	-
This work	ADFCNN	**86.87**	**87.26**	**79.39**	**87.81**

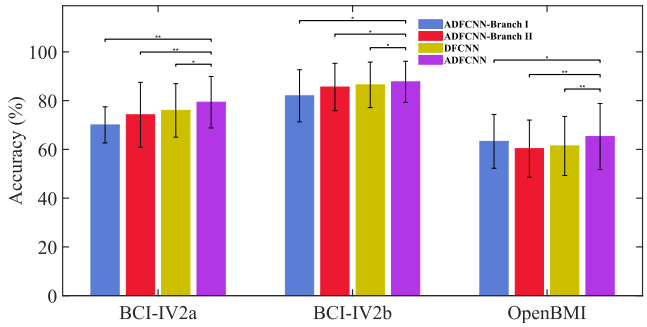

Fig. 6. The ablation experimental results, * : $p < 0.05$, ** : $p < 0.01$..

- ADFCNN-Branch I: This variant model mainly consists of a large-scale temporal convolution layer, separable convolution layers, an average pooling layer for feature extraction, and a dense layer for feature classification, all of which were used to demonstrate the effectiveness of the first temporal-spatial CNN branch.
- ADFCNN-Branch II: To demonstrate the effectiveness of the second temporal-spatial CNN branch, this variant model consists of a small-scale temporal convolution layer, a standard spatial convolution layer, an average pooling layer for feature extraction, and a dense layer for feature classification.
- DFCNN: This model was designed to demonstrate the effectiveness of the self-attention mechanism in the feature fusion module.

Fig. 6 demonstrates that the critical modules of the proposed method are successful in MI classification, specifically temporal convolution, separable spatial convolution, standard spatial convolution and self-attention layers. It can be observed that the dual-scale joint temporal-spatial CNN yields an improved classification performance compared to the single-branch temporal-spatial CNN across three datasets. Furthermore, the self-attention mechanism can be found to be beneficial, leading to a further increase in performance.

D. Convolution Kernel Visualization

To further understand the effect of convolution kernels in the proposed method, we visualize the weights of dual-scale spatial and temporal convolution kernels in ADFCNN by topological mapping and Fourier transform, respectively. Fig. 7(a) depicts the topological mapping of weights from separable spatial kernels, revealing that these eight kernels mainly extract global spatial information from all electrodes. Fig. 7(b) shows the topological mapping of weights from sixteen standard spatial convolution kernels. We observe that the weights of MI-related electrodes such as C3, C4 and Cz are higher than that of other electrodes in most kernels, revealing that they have learned more detailed spatial information related to MI. Additionally, the Fourier transform visualization of weights from dual-scale temporal convolution kernels further shows evidence of the efficacy of the proposed method. Fig. 8 reveals that temporal convolution kernels maintain specific spectral information, with each kernel curve exhibiting a single peak in the power spectral density (PSD). In Fig. 8(a), five kernels are observed to learn spectral information at approximately 10 Hz, and three kernels are present to retain the frequency information in the 20-40 Hz range. These findings demonstrate that large-scale temporal convolution kernels with size [1 × 125] can broadly capture low-frequency information from EEG μ rhythm. Furthermore, Fig. 8(b) indicates that most temporal kernels maintain frequency information between 20 Hz and 30 Hz, with only a few kernels retaining frequency information in the 10-20 Hz range. These results suggest that small-scale temporal convolution kernels of size [1×30] can primarily capture high-frequency information from EEG β rhythm.

E. Attention Visualization

For a detailed interpretation of the self-attention mechanism applied in feature fusion, we visualize the self-attention matrix to analyze the internal similarity of concatenated features. Fig. 9 shows the self-attention matrix of concatenated features

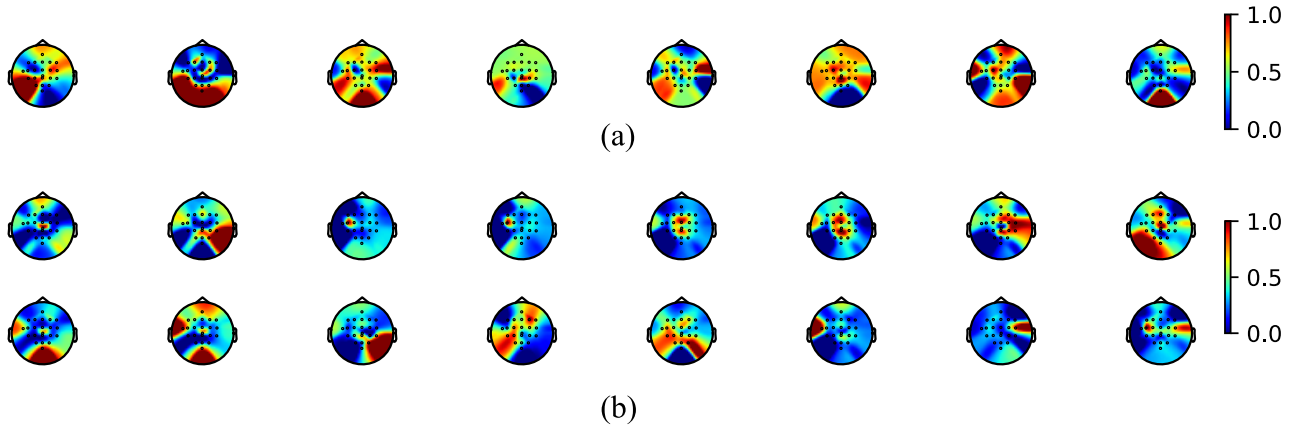

Fig. 7. The spatial visualization of spatial convolution kernels after learning: (a) eight separable spatial convolution kernels in Branch I; (b) sixteen standard spatial convolution kernels in Branch II (Subject 7 from BCI-IV2a dataset).

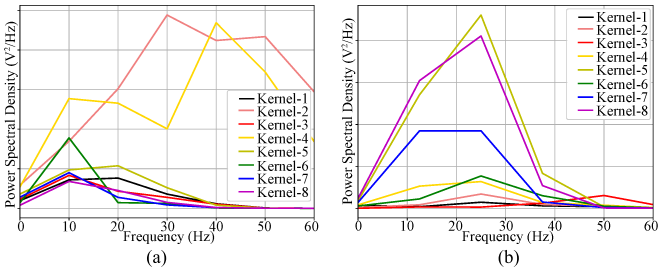

Fig. 8. The spectral visualization of temporal convolution kernels after learning: (a) kernel with size 1×125 in Branch I; (b) kernel with size 1×30 in Branch II (Subject 7 from BCI-IV2a dataset).

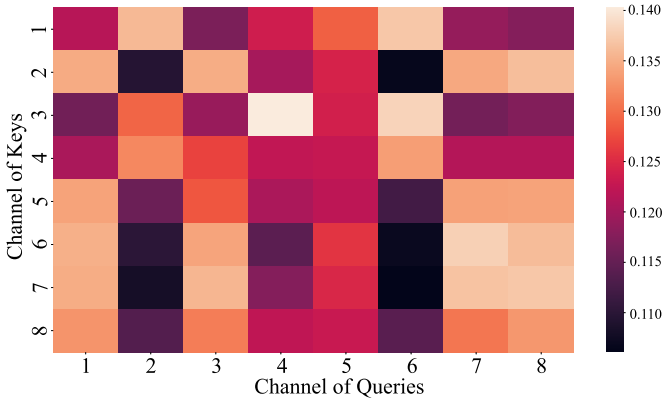

Fig. 9. The self-attention score visualization (Subject 7 from BCI-IV2a dataset).

from the subject 7 in BCI-IV2a. According to Equation 13, the self-attention score is calculated between Queries and Keys from concatenated features through a dot-product and a softmax operation, which reflects the similarity within concatenated features. As can be observed in Fig. 9, channels 1, 3, 7, and 8 of the Queries feature map are more similar to channels 5, 6, 7 and 8 of the Keys feature map for this subject, which are allocated more self-attention scores. The self-attention scores can then be used to reweigh the Values feature map, thereby adaptively improving the flexibility of fusion feature.

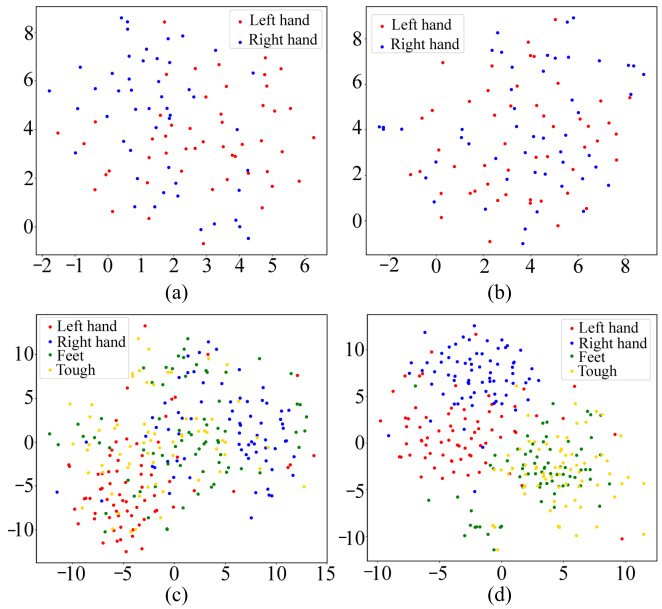

Fig. 10. The t-SNE visualization for high-dimensional features from two branches of ADFCNN: (a) the two-class distribution of features by Branch I for subject 7 from OpenBMI dataset, (b) the two-class distribution of features by Branch II for subject 7 from OpenBMI dataset, (c) the four-class distribution of features by Branch I for subject 7 from BCI-IV2a dataset, (d) the four-class distribution of features by Branch II for subject 7 from BCI-IV2a dataset.

F. Feature Visualization and Confusion Matrices

In order to further illustrate the efficacy of the two temporal-spatial CNN branches and feature fusion based on the self-attention mechanism, the t-distributed Stochastic Neighbor Embedding (t-SNE) method is applied to visualize the high dimensional features extracted by each of the four models, Branch I, Branch II, DFCNN, and ADFCNN. As seen in Fig. 10, two temporal-spatial CNN branches are found to have various learning abilities of features. Specifically, the distribution of features learned by Branch I (Fig. 10(a)) is more separable than that by Branch II (Fig. 10(b)). In contrast, the distribution of features learned by Branch II (Fig. 10(d)) is more separable than that by Branch I (Fig. 10(c)). Furthermore, Fig. 11 demonstrates that the features with self-attention

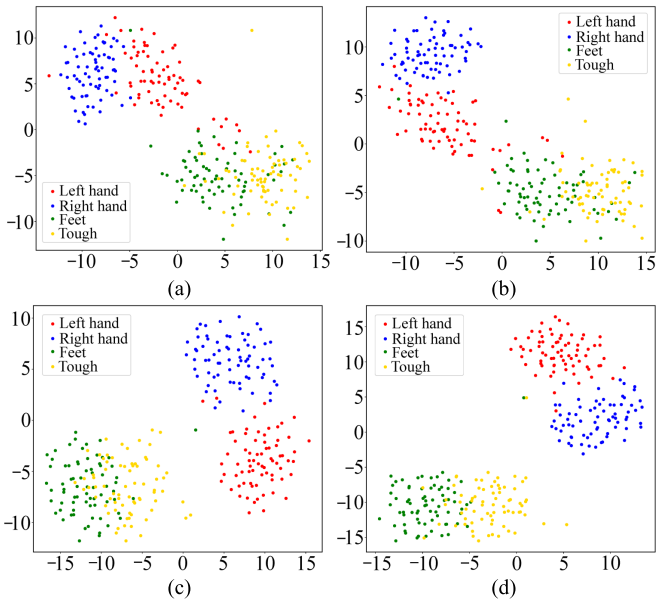

Fig. 11. The t-SNE visualization for high-dimensional features from concatenation fusion and attention-based fusion: (a) the four-class distribution of fusion features by concatenation for subject 1 from BCI-IV2a dataset, (b) the four-class distribution of fusion features by attention for subject 1 from BCI-IV2a dataset, (c) the four-class distribution of fusion features by concatenation for subject 7 from BCI-IV2a dataset, (d) the four-class distribution of fusion features by attention for subject 7 from BCI-IV2a dataset.

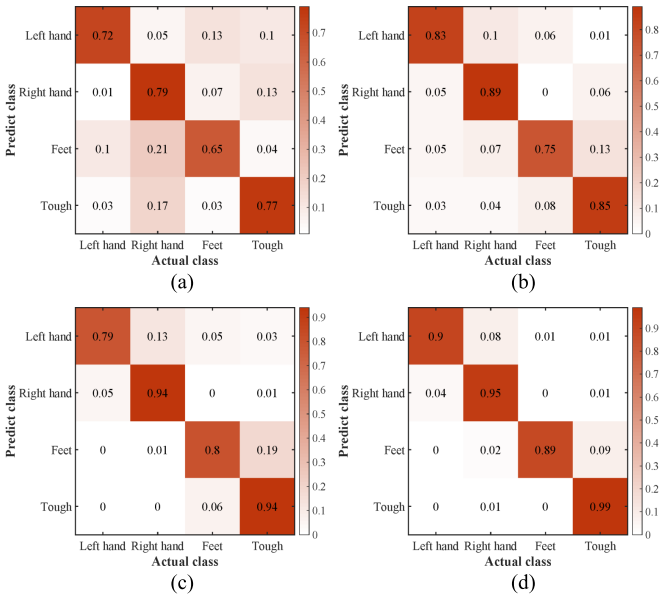

Fig. 12. Confusion matrices of classification accuracy for each class in the testing data of subject 7 from BCI-IV2a dataset: (a) confusion matrix of ADFCNN-Branch I, (b) confusion matrix of ADFCNN-Branch II, (c) confusion matrix of DFCNN, (d) confusion matrix of ADFCNN.

fusion are more separable than those without the self-attention mechanism. For instance, the clusters of left-hand and right-hand are separated but there is still overlap of feet and tough clusters as seen in Figs. 11(a) and 11(c). However, with the aid of self-attention processing, the overlap between feet and tough clusters has been alleviated, as seen in Figs. 11(b) and 11(d). These results show that the self-attention mechanism can effectively improve the flexibility of concatenated features and contribute to the attainment of discriminative and robust fusion features.

Moreover, Fig. 12 shows the confusion matrix for each class with the proposed model and three variants as introduced in ablation experiments. We find that the classification accuracy of ADFCNN is 0.93, significantly higher than 0.73, 0.83 and 0.87 of ADFCNN-Branch I, ADFCNN-Branch II and DFCNN, respectively, for subject 7. Compared with DFCNN, the ADFCNN has an improved true positive rate for Feet and Left-hand as shown in Figs. 12 (c) and 12(d), which is consistent with corresponding feature distributions as shown in Figs. 11(c) and 11(d).

V. Discussion

Recent studies have explored the use of convolutional neural networks (CNNs) for motor-imagery (MI) classification, including single-scale and multi-scale CNNs. Single-scale CNNs typically utilize temporal and spatial convolution with a single-scale to extract spectral and spatial information from electroencephalography (EEG) signals. For example, temporal convolutions with different kernel sizes capture different-band spectral information, and spatial convolutions with separable or standard manners capture different inter-channel information. Multi-scale CNNs capture features from different scales, but traditional feature fusion of multi-scale CNNs only involves direct feature concatenation. While this simple concatenation increases the dimension of features and thus the amount of information, it is limited in its ability to explore implicit information of the fused features. To address this, we propose an attention-based dual-scale fusion CNN (ADFCNN) for MI-based BCI. Specifically, the dual-scale temporal-spatial CNN jointly extracts abundant spectral and spatial information, and the self-attention mechanism performs feature fusion according to the internal similarity of the concatenated features. By utilizing similar but different structures in Branch-I and Branch-II, dual-scale joint CNN can effectively extract multi-scale spectral and spatial information from the input signals. This design choice enhances the overall performance of our proposed method by optimizing the extraction of diverse and complementary features. Additionally, this scheme also provides novel insight through self-attention for effective information fusion from different scales. Moreover, We conducted ablation experiments to explain the mechanism and demonstrate the effectiveness of our approach, and used visualization techniques to provide interpretability into the learning process and feature distributions of our model.

In our experiments, we focus on processing the EEG signals within the post-cue period of 0-3 seconds and a frequency band of 0-40 Hz. This choice aligns with the deep learning methods employed in previous studies [27], [53], [56] and proves to be effective in capturing a broader range of motor-related patterns by a series of experiments as shown in Tables S1-S4 of the supplementary file. By incorporating this time period and frequency band selection, deep models successfully encompass both the low-frequency information provided by

the motor-related cortical potential (MRCP) [57] and the high-frequency information derived from the sensory motor rhythm (SMR) [58]. As a result, deep learning methods can demonstrate improved performances when operating within this particular period and frequency band selection. Moreover, we compare the proposed method to five representative deep learning baselines, four SOTA multi-scale CNNs methods and three latest deep learning techniques, ultimately achieve superior classification performances on three public datasets for subject-specific MI-based BCI. Through an ablation study, we demonstrate the effectiveness of dual-scale joint temporal-spatial CNN and self-attention mechanism modules. Additionally, we explore the effect of critical parameters in the model and find that the number of temporal convolution kernels influences the model performance with limited training data. To reveal the learning process and learned feature distribution of the proposed method, we perform detailed visualizations. Firstly, spatial convolution visualization reveals that separable spatial kernels can mainly extract the global information from all electrodes, while standard spatial convolution kernels can extract more detailed information related to MI by concentrating higher weights in brain central areas including C3, C4 and Cz electrodes. Secondly, temporal convolution visualization reveals that the low-frequency information extracted by large-scale temporal kernels is mainly located in the μ rhythm (7-13 Hz), and high-frequency information extracted by small-scale temporal kernels is mainly located in the β rhythm (13-30 Hz). These rhythms can reflect the event-related desynchronization (ERD), event-related synchronization (ERS) patterns in motor imagery [58]. Thirdly, the self-attention visualization can reflect the internal similarity of concatenated features and be used to generate more flexible fusion features. Finally, by visualizing features with t-SNE, we observe that two temporal-spatial CNN branches have different learning abilities on the same subject. This demonstrates that it is more reasonable to design a dual-scale joint CNN module to capture different-scale information compared to a single-scale CNN. Additionally, we observe that self-attention mechanism can further enhance the discrimination capability of learned features and improve the feature fusion in multi-scale CNN.

There are still some limitations in this work. First, due to the compact nature of the proposed network and its fewer parameters, data augmentation techniques have not been considered to increase the training dataset. Second, the primary focus has been the discussion of the critical parameter of temporal kernel number, while other network parameters such as temporal convolution kernel size and pooling size have been taken from existing studies. Third, while current study focuses on subject-specific task, the proposed neural network structure can be applied as a feature extractor in cross-subject task as well, and we intend to explore the applicability and adaptability of our method in cross-subject scenarios in future research.

VI. Conclusion

In this paper, we present a highly effective approach to MI classification using an attention-based dual-scale fusion convolutional neural network (ADFCNN). To extract more abundant information over a wide spectrum from EEG signals and effectively fuse information from different scales, our method jointly extracts EEG spectral and spatial information at different scales, while also employing a self-attention mechanism to enhance the flexibility of feature fusion. Empirical results on three public datasets demonstrate that our method outperforms several state-of-the-art methods for subject-specific MI-based BCI. Further validations of the proposed method are provided by ablation experiments, which demonstrate the mechanisms of feature extraction and fusion, and visualization analyses, which reveal the characteristics of different-scale temporal-spatial convolutions and the self-attention mechanism. To sum up, the proposed ADFCNN yields promising results for MI-based BCI and provides novel insight through self-attention for different-scale information fusion, is thus an effective feature extractor for EEG processing with deep learning.

References

[1] P. Chen, H. Wang, X. Sun, H. Li, C. Grebogi, and Z. Gao, "Transfer learning with optimal transportation and frequency mixup for EEG-based motor imagery recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2866–2875, 2022, doi: 10.1109/TNSRE.2022.3211881.

[2] Q. She, T. Chen, F. Fang, J. Zhang, Y. Gao, and Y. Zhang, "Improved domain adaptation network based on Wasserstein distance for motor imagery EEG classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1137–1148, 2023, doi: 10.1109/TNSRE.2023.3241846.

[3] F. Wei, X. Xu, T. Jia, D. Zhang, and X. Wu, "A multi-source transfer joint matching method for inter-subject motor imagery decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1258–1267, 2023, doi: 10.1109/TNSRE.2023.3243257.

[4] X. Tang, C. Yang, X. Sun, M. Zou, and H. Wang, "Motor imagery EEG decoding based on multi-scale hybrid networks and feature enhancement," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1208–1218, 2023, doi: 10.1109/TNSRE.2023.3242280.

[5] X. Liu et al., "Activation network improves spatiotemporal modelling of human brain communication processes," *NeuroImage*, vol. 285, Jan. 2024, Art. no. 120472.

[6] C. M. Wong, B. Wang, Z. Wang, K. F. Lao, A. Rosa, and F. Wan, "Spatial filtering in SSVEP-based BCIs: Unified framework and new improvements," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 11, pp. 3057–3072, Nov. 2020.

[7] B. Abibullaev and A. Zollanvari, "A systematic deep learning model selection for P300-based brain–computer interfaces," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 52, no. 5, pp. 2744–2756, May 2022.

[8] W. Tao et al., "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 382–393, Jan. 2023.

[9] B. Wang et al., "Common spatial pattern reformulated for regularizations in brain–computer interfaces," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 5008–5020, Oct. 2021.

[10] A. Al-Saegh, S. A. Dawwd, and J. M. Abdul-Jabbar, "Deep learning for motor imagery EEG-based classification: A review," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102172.

[11] H. Altaheri et al., "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14681–14722, Jul. 2023.

[12] P. Chen, Z. Gao, M. Yin, J. Wu, K. Ma, and C. Grebogi, "Multiattention adaptation network for motor imagery recognition," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 52, no. 8, pp. 5127–5139, Aug. 2022.

[13] J. Fumanal-Idocin, Y.-K. Wang, C.-T. Lin, J. Fernández, J. A. Sanz, and H. Bustince, "Motor-imagery-based brain–computer interface using signal derivation and aggregation functions," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7944–7955, Aug. 2022.

[14] H. Raza, A. Chowdhury, and S. Bhattacharyya, "Deep learning based prediction of EEG motor imagery of stroke patients' for neuro-rehabilitation application," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[15] J. Long, Y. Li, H. Wang, T. Yu, J. Pan, and F. Li, "A hybrid brain computer interface to control the direction and speed of a simulated or real wheelchair," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 5, pp. 720–729, Sep. 2012.

[16] K. K. Ang and C. Guan, "EEG-based strategies to detect motor imagery for control and rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 4, pp. 392–401, Apr. 2017.

[17] S. Aggarwal and N. Chugh, "Signal processing techniques for motor imagery brain computer interface: A review," *Array*, vols. 1–2, Jan. 2019, Art. no. 100003.

[18] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains," *Int. Scholarly Res. Notices*, vol. 2014, pp. 1–7, Feb. 2014.

[19] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clin. Neurophysiol.*, vol. 110, no. 5, pp. 787–798, May 1999.

[20] Y. Wang, K. C. Veluvolu, and M. Lee, "Time-frequency analysis of band-limited EEG with BMFLC and Kalman filter for BCI applications," *J. NeuroEng. Rehabil.*, vol. 10, no. 1, pp. 1–16, Dec. 2013.

[21] K. Samiee, P. Kovács, and M. Gabbouj, "Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 541–552, Feb. 2015.

[22] M. Bentlemsan, E.-T. Zemouri, D. Bouchaffra, B. Yahya-Zoubir, and K. Ferroudji, "Random forest and filter bank common spatial patterns for EEG-based motor imagery classification," in *Proc. 5th Int. Conf. Intell. Syst., Modeling Simulation*, Jan. 2014, pp. 235–238.

[23] L. Quoc Thang and C. Temiyasathit, "Increase performance of four-class classification for motor-imagery based brain-computer interface," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2014, pp. 1–5.

[24] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "An empirical mode decomposition based filtering method for classification of motor-imagery EEG signals for enhancing brain-computer interface," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–7.

[25] S. Bhattacharyya, A. Khasnobish, S. Chatterjee, A. Konar, and D. N. Tibarewala, "Performance analysis of LDA, QDA and KNN algorithms in left-right limb movement classification from EEG data," in *Proc. Int. Conf. Syst. Med. Biol.*, Dec. 2010, pp. 126–131.

[26] X. Chen, C. Li, A. Liu, M. J. McKeown, R. Qian, and Z. J. Wang, "Toward open-world electroencephalogram decoding via deep learning: A comprehensive survey," *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 117–134, Mar. 2022.

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

[28] 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.

[29] D. Milanés Hermosilla et al., "Shallow convolutional network excel for classifying motor imagery EEG in BCI applications," *IEEE Access*, vol. 9, pp. 98275–98286, 2021.

[30] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "A multi-view CNN with novel variance layer for motor imagery brain computer interface," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 2950–2953.

[31] J. Wang, L. Yao, and Y. Wang, "IFNet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1900–1911, 2023.

[32] H. K. Lee and Y.-S. Choi, "A convolution neural networks scheme for classification of motor imagery EEG based on wavelet time-frequecy image," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2018, pp. 906–909.

[33] G. Dai, J. Zhou, J. Huang, and N. Wang, "HS-CNN: A CNN with hybrid convolution scale for EEG motor imagery classification," *J. Neural Eng.*, vol. 17, no. 1, Jan. 2020, Art. no. 016025.

[34] W. Ko, E. Jeon, S. Jeong, and H.-I. Suk, "Multi-scale neural network for EEG representation learning in BCI," *IEEE Comput. Intell. Mag.*, vol. 16, no. 2, pp. 31–45, May 2021.

[35] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3D convolutional neural network for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2164–2177, Oct. 2019.

[36] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[37] M. X. Cohen, *Analyzing Neural Time Series Data: Theory and Practice*. Cambridge, MA, USA: MIT Press, 2014.

[38] Z. Jia, Y. Lin, J. Wang, K. Yang, T. Liu, and X. Zhang, "MMCNN: A multi-branch multi-scale convolutional neural network for motor imagery classification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2021, pp. 736–751.

[39] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[40] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6298–6306.

[41] C. Yan et al., "STAT: Spatial–temporal attention mechanism for video captioning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 229–241, Jan. 2020.

[42] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent EEG signal analysis," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 715–719, May 2019.

[43] 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.

[44] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008–Graz data set A," Inst. Knowl. Discovery (Lab. Brain-Comput. Interfaces), Graz Univ. Technol., Graz, Austria, Tech. Rep., 2008, pp. 1–6, vol. 16.

[45] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008–Graz data set B," Graz Univ. Technol., Austria, Graz, Austria, Tech. Rep., 2008, pp. 1–6.

[46] M.-H. Lee et al., "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, May 2019, Art. no. giz002.

[47] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Gener. Comput. Syst.*, vol. 101, pp. 542–554, Dec. 2019.

[48] X. Liu, Y. Shen, J. Liu, J. Yang, P. Xiong, and F. Lin, "Parallel spatial–temporal self-attention CNN-based motor imagery classification for BCI," *Frontiers Neurosci.*, vol. 14, Dec. 2020, Art. no. 587520.

[49] M. Riyad, M. Khalil, and A. Adib, "MI-EEGNET: A novel convolutional neural network for motor imagery classification," *J. Neurosci. Methods*, vol. 353, Apr. 2021, Art. no. 109037.

[50] X. Tang, C. Yang, X. Sun, M. Zou, and H. Wang, "Motor imagery EEG decoding based on multi-scale hybrid networks and feature enhancement," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1208–1218, 2023, doi: 10.1109/TNSRE.2023.3242280.

[51] C. Ju and C. Guan, "Tensor-CSPNet: A novel geometric deep learning framework for motor imagery classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10955–10969, Dec. 2023, doi: 10.1109/TNNLS.2022.3172108.

[52] C. Liu et al., "SincNet-based hybrid neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 540–549, 2022.

[53] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2023, doi: 10.1109/TNSRE.2022.3230250.

[54] H.-J. Hwang, K. Kwon, and C.-H. Im, "Neurofeedback-based motor imagery training for brain–computer interface (BCI)," *J. Neurosci. Methods*, vol. 179, no. 1, pp. 150–156, Apr. 2009.

[55] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics*. Cham, Switzerland: Springer, 1992, pp. 196–202.

[56] S. U. Amin, H. Altaheri, G. Muhammad, W. Abdul, and M. Alsulaiman, "Attention-inception and long-short-term memory-based electroencephalography classification for motor imagery tasks in rehabilitation," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5412–5421, Aug. 2022.

[57] T. Liu, G. Huang, N. Jiang, L. Yao, and Z. Zhang, "Reduce brain computer interface inefficiency by combining sensory motor rhythm and movement-related cortical potential features," *J. Neural Eng.*, vol. 17, no. 3, Jun. 2020, Art. no. 035003.

[58] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. L. da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, May 2006.