A Multi-Domain Convolutional Neural Network for EEG-Based Motor Imagery Decoding

Hongyi Zhi, Zhuliang Yu, *Member, IEEE*, Tianyou Yu, *Member, IEEE*, Zhenghui Gu, and Jian Yang

***Abstract*— Motor imagery (MI) decoding plays a crucial role in the advancement of electroencephalography (EEG)-based brain-computer interface (BCI) technology. Currently, most researches focus on complex deep learning structures for MI decoding. The growing complexity of networks may result in overfitting and lead to inaccurate decoding outcomes due to the redundant information. To address this limitation and make full use of the multi-domain EEG features, a multi-domain temporal-spatial-frequency convolutional neural network (TSFCNet) is proposed for MI decoding. The proposed network provides a novel mechanism that utilize the spatial and temporal EEG features combined with frequency and time-frequency characteristics. This network enables powerful feature extraction without complicated network structure. Specifically, the TSFCNet first employs the MixConv-Residual block to extract multi-scale temporal features from multi-band filtered EEG data. Next, the temporal-spatial-frequency convolution block implements three shallow, parallel and independent convolutional operations in spatial, frequency and time-frequency domain, and captures high discriminative representations from these domains respectively. Finally, these features are effectively aggregated by average pooling layers and variance layers, and the network is trained with the joint supervision of the cross-entropy and the center loss. Our experimental results show that the TSFCNet outperforms the state-of-the-art models with superior classification accuracy and kappa values (82.72% and 0.7695 for dataset BCI competition IV 2a, 86.39% and 0.7324 for dataset BCI competition IV 2b). These competitive results demonstrate that the proposed network is promising for enhancing the decoding performance of MI BCIs.**

***Index Terms*— Brain–computer interface (BCI), electroencephalography (EEG), motor imagery (MI), convolutional neural network (CNN), center loss.**

I. Introduction

BRAIN-COMPUTER interface (BCI), an advanced external information exchange and control technology, is able to directly connect human brain and other electronic devices without the involvement of peripheral nerves and muscles [1]. This technology has broad applications in the field of rehabilitation medicine [2] [3]. Among BCI systems, motor imagery (MI) is one of the most popular electroencephalography (EEG)-based paradigms, which can trigger neuronal activities in the sensorimotor network of specific brain areas similar to the way as the real physical movement [4].

During the MI process, the rhythmic EEG activity is suppressed or enhanced in the sensorimotor area of the contralateral hemisphere and the ipsilateral hemisphere, respectively. The former case is known as event-related desynchronization (ERD), and the latter one is event-related synchronization (ERS) [5]. By decoding the ERD/ERS in the EEG correctly, the people with severe motor diseases can control external devices via movement intents. Therefore, the classification of EEG signals plays a crucial role in the research of MI BCIs and requires advanced signal decoding techniques.

EEG-based MI decoding for BCI classification encounters several significant challenges, e.g., the low signal-to-noise ratio, high intertrial variability and sensitivity to noise [6]. Previous studies on MI decoding can be broadly classified into two categories: classical machine learning methods and deep learning methods. Among the classical machine learning methods, Common Spatial Patterns (CSP) is one of the most powerful methods constructing optimal spatial filters [7]. Consequently, a large number of extended CSP variants have emerged such as the filter bank CSP (FBCSP) [8] and the discriminative filter bank CSP (DFBCSP) [9]. For the feature classification, many classical classifiers, such as support vector machines (SVMs) and linear discriminant analysis (LDA) are applied to classify the MI discriminative features.

These approaches rely heavily on handcrafted features and therefore suffer from several drawbacks, including time-consuming, subject-dependent and poor feature extraction capabilities. Manually designed features may lead to poor decoding performance of MI-EEG due to the limitations of human knowledge and experience. Additionally, the ideas of optimal frequency band and spatial filter selection fail to address the issue of heterogeneity among subjects, thus lacking diversity.

Recently, deep learning, as an extensively data-driven method, has achieved state-of-the-art (SOTA) performance in the EEG classification task and gained successes in addressing the aforementioned challenges [10] [11]. The convolutional

Manuscript received 4 July 2023; revised 7 September 2023; accepted 5 October 2023. Date of publication 10 October 2023; date of current version 18 October 2023. This work was supported in part by STI2030-Major Projects under Grant 2022ZD0211700; and in part by the National Natural Science Foundation of China under Grant 61836003, Grant 61906211, Grant 62376098, and Grant 62276102. *(Corresponding author: Jian Yang.)*

The authors are with the School of Automation Science and Engineering, South China University of Technology, Guangdong 510641, China, and also with the Pazhou Laboratory, Guangzhou 510641, China (e-mail: auzhihongyi@mail.scut.edu.cn; zlyu@scut.edu.cn; auyuty@scut.edu.cn; zhgu@scut.edu.cn; yangjianxin@scut.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3323325

neural network (CNN) based deep learning architectures employ one or several customized kernel matrices to extract hybrid features from the raw data. Schirrmeister et al. proposed the deep ConvNet and showed the potential of CNN architecture for EEG decoding [12]. EEGNet is another widespread method proposed by Lawhern et al., which can extract temporal and spatial features simultaneously [13]. In [14], Li et al. proposed a novel multi-layers 1D-CNN neural network architecture called CP-MixedNet. To address the issue that the convolutional kernel size is generally fixed, the study in [15] proposed a CNN with hybrid convolution scales. Similarly, the EEG-Inception proposed by Zhang et al. uses several inceptions and residual modules as the backbone with high potential for the subject-independent EEG-based MI classification [16]. Furthermore, the multi-view methods achieved promising results. For instance, FBCNet and FBMSNet both implement temporal-spatial convolution to filtered EEG data [17] [18]. A recent benchmark network, namely EEGNeX, is a pure convolution-based architecture derived from analogy investigations between the EEG and neural network architecture [19]. Besides, Altaheri et al. proposed attention-based temporal convolutional network ATCNet and D-ATCNet and validated them on BCI competition IV 2a dataset [20] [21].

However, the growing complexity of CNNs may result in overfitting and lead to inaccurate decoding outcomes due to the marginal effect and the presence of redundant information. Moreover, these methods obtain deep features only from temporal and spatial domains. It limits their capabilities to develop highly distinguishable feature representations. What is more, recent works [22] [23] indicate that the conventional cross-entropy (CE) loss is ineffective in reducing intraclass variation, which may cause the poor performance on EEG classification.

To tackle the issues stated above, in this study, a multi-domain temporal-spatial-frequency convolutional neural network (TSFCNet) is proposed for MI-EEG decoding. Specifically, the MixConv-Residual block is first employed to extract multiscale temporal features from the multi-band filtered EEG data, followed by residual connections. The temporal-spatial-frequency convolution (TSF-Conv) block is then designed to implement three parallel and independent convolutions in spatial, frequency and time-frequency domains for capturing highly discriminative multi-domain features respectively. Moreover, inspired by [22], we apply the center loss as auxiliary costs for the proposed framework to increase the discrimination of different classes of samples in the feature space. Meanwhile, the center loss can minimize the distances between the learned representations and the centers of their corresponding classes. Finally, with the joint supervision of the CE loss and the center loss, these three feature representations are effectively aggregated by average pooling layer and variance layer, and a fully connected (FC) layer is used for classification. The proposed TSFCNet is evaluated on three public BCI datasets, and ablation experiments are also conducted to demonstrate the effectiveness of each module used in the proposed TSFCNet method.

The major contributions of this article are summarized as follows.

1) A multi-domain framework named TSFCNet is proposed for MI decoding, which is able to effectively capture highly discriminative and robust features with three shallow, parallel and independent convolutions. It enables powerful feature extraction without complicated network structure.
2) The proposed TSFCNet with MixConv-Residual block and TSF-Conv block provides a novel mechanism that leverages spatial and temporal EEG features combined with frequency and time-frequency characteristics to improve the EEG decoding.
3) Numerical experiment results show that the TSFCNet outperforms the SOTA methods. The extensive ablation studies validate the effectiveness of each block in the TSFCNet.

The rest of our paper is organized as follows. Section II details the proposed TSFCNet method. Section III performs numerical experiments and extensive ablation studies, and presents the experimental results. Finally, Section IV and section V present discussions and conclusions in this paper, respectively.

II. Methodology

In this section, the preprocessing step is first described. Then the basic blocks of the TSFCNet are introduced, including the MixConv-Residual block, the TSF-Conv block and the classifier. Finally, the loss functions of the TSFCNet are introduced. The source code of the proposed method is available at https://github.com/hongyizhi/TSFCNet.

A. Preprocessing

Consider a set of single-trial raw EEG data that are as $X_n = \{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^{C\times T}$ with corresponding label set $Y_n = \{y_i\}_{i=1}^n$, where n represents the number of EEG trials, C represents the number of EEG channels and T represents the time points.

The filtered EEG signals $X_{FB} \in \mathbb{R}^{N_f\times C\times T}$ are generated by Chebyshev Type II bandpass filters based on the predefined frequency filter bands $F = \{f_i\}_{i=1}^{N_f}$, where N_f is the number of filter bands. Earlier work [24] has already shown that mu (8-12 Hz), beta (12-32 Hz) and also theta (4-7 Hz) frequency bands play crucial roles in MI tasks. Therefore, we construct the specified filter F by using 9 nonoverlapping frequency bands, each with a 4 Hz bandwidth. The frequency filter bands F spanning from 4 to 40 Hz (i.e., 4-8, 8-12..., 36-40 Hz). Following the given certain frequency filter banks F, the filtered EEG signals X_{FB} are deterministically obtained as follows:

$$X_{FB} = F \otimes X_n \tag{1}$$

where, $\otimes$ indicates bandpass filtering operation.

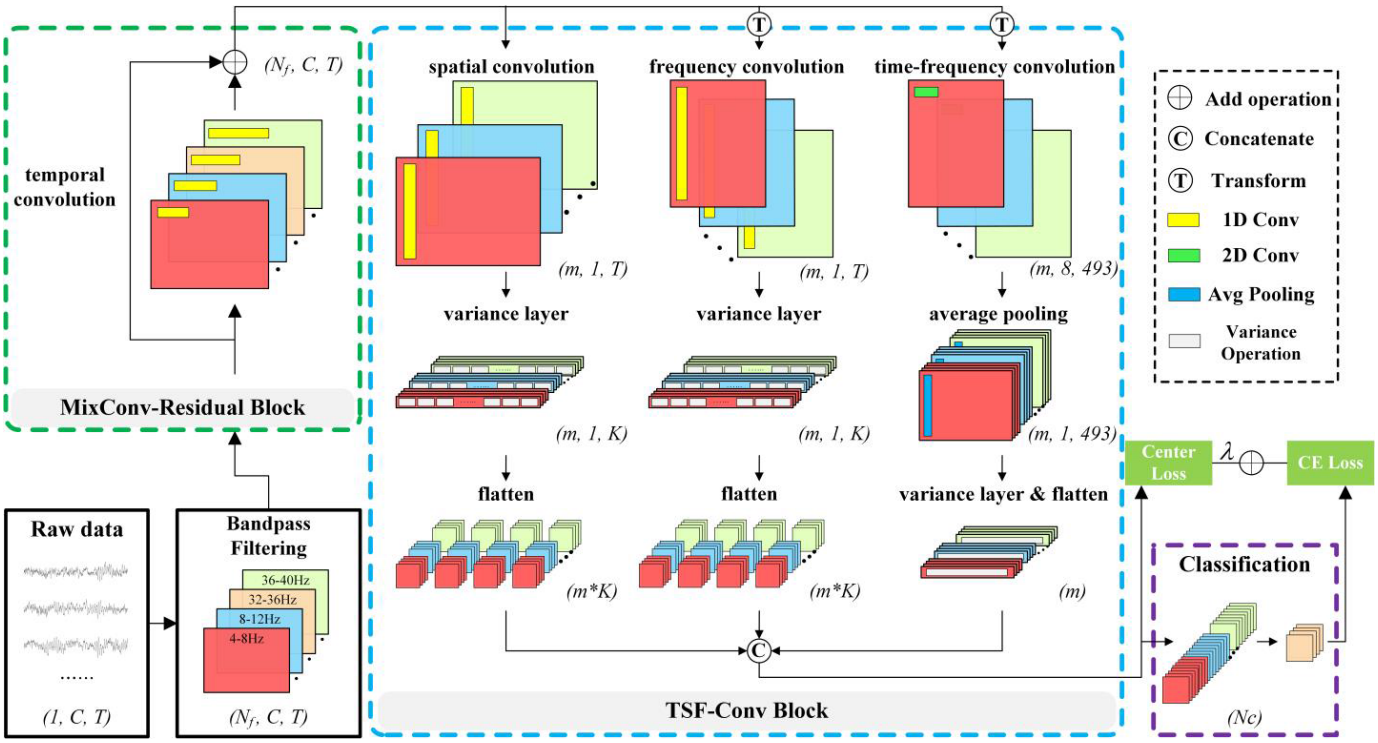

Fig. 1. Architecture of the proposed TSFCNet. N_f, C and T represent the number of filter bands, the number of EEG channels and the number of time points, respectively. m, K, N_c and λ donate the number of kernels, total number of nonoverlapping windows, the number of output classes and the trade-off scalar, respectively.

B. Temporal-Spatial-Frequency Convolutional Neural Network

In this section, we describe the proposed TSFCNet in detail. The TSFCNet consists of the MixConv-Residual block, the TSF-Conv block and the classifier. The overall structure of the proposed TSFCNet is depicted in Fig. 1.

1) Design of the MixConv-Residual Block: The MixConv-Residual block is designed based on two novel ideas that take advantage of the temporal characteristics of the EEG. The first idea is to implement mixed depthwise convolution (MixConv) to extract multiscale temporal information from the filtered EEG data [25]. The MixConv mixes up multiple kernel sizes in a single convolution without changing the macro-architecture of the neural networks which improves the accuracy and efficiency of convolution network. Note that a single convolution scale would lead to the limited classification accuracy, since the optimal scale may be distinct among different subjects, or at different times for the same subject. Thus, the MixConv is a solution to achieve wider and multiscale feature extraction with strong robustness to subject-dependency.

He et al. [26] effectively solved the learning degradation problem by applying the residual learning framework. The second idea known as residual connections takes inspiration from the ResNet. By simply driving the residual of the multiple nonlinear layers toward zero, the residual connections guarantee to approach the identity mappings. Meanwhile, shortcuts are straightforward implementations of identity mappings. It provides a path for the information to flow unmodified through the whole architecture.

Specifically, in the MixConv layer, N_f 1D convolutional kernels with the sizes of (1, 15), (1, 31), (1, 63) and (1, 125) are used to learn the temporal features. The number of kernels is equal to that of filter banks so that we can easily implement residual connections for the outputs of the MixConv layer. The outputs of the MixConv-Residual block are defined as $X_{MixConv-residual} \in \mathbb{R}^{N_f \times C \times T}$. It includes the filtered EEG information and the temporal features extracted from the MixConv-Residual Block. The outputs are then fed into the TSF-Conv block for further feature extraction.

2) Design of the TSF-Conv block: The TSF-Conv block is composed of three parallel and independent convolutional operations. Each convolutional operation includes a convolutional layer, a batch normalization (BN) layer, an activate layer and a variance layer. These convolutional operations learn the spatial, frequency and time-frequency features respectively.

The tensor flows involved in these three convolutional operations are as follows. For the spatial convolutional layer, m kernels of size $(C, 1)$ are used to fuse the spatial information from all input channels, where m is set to be 36 for 4 four-classes dataset and 9 for two-classes dataset. This operation fuses the spatial information to the features of a single channel. For the frequency convolution layer, the shape of the input tensor $X_{MixConv-residual}$ is first permutated from $N_f \times C \times T$ to $C \times N_f \times T$ by transformation operation. It is similar to the spatial convolution layer, the frequency convolution layer with m kernels of size $(N_f, 1)$ mixes the frequency feature from all different filter banks. Also, the convolution kernel is designed to span across all the frequency filter banks, and reduce its frequency dimension to 1. In the time-frequency

convolution layer, after following the same transformation operation as described above, we apply m small 2D kernels with the size of (2, 16) to obtain the frequency feature along the time. The reason we choose a small kernel size is to reduce the parameters. This layer implicitly transforms the tensor into a combination of temporal and frequency representation and enhances the feature extraction of EEG signals. Besides, BN layer is adopted after each convolutional layer to alleviate the overfitting problem and tune the optimal parameters of the neural networks. The exponential linear unit (ELU) is employed as the activation function to overcome the vanishing gradient problem. As a result, TSF-Conv outputs three feature maps x_{conv} with shape of $m \times D \times \tilde{T}$, where D denotes the dimension of the output of the TSF-Conv, and $\tilde{T}$ is the time dimension.

The output data of TSF-Conv still contain a large amount of information along the time dimension. It requires further processing. As in [17], instead of maximum pooling and average pooling, we apply the variance operations to efficiently extract the most relevant temporal features. Such kind of variance layer considers the differences of various classes in their spectral power (ERD/ERS). And the variance layer thus becomes a more suitable option for EEG temporal characterization [17]. The variance layer is defined as:

$$x_{var}(m,\, d,\, k) = \frac{K}{T} \sum_{t=k*\omega}^{(k+1)*\omega-1} (x_{conv}(m,\, d,\, t) - \mu(m,\, d,\, k))^2 \tag{2}$$

where $x_{var}(m,\, d,\, k)$ and $\mu(m,\, d,\, k)$ are the variance layer result and the temporal mean of $x_{conv}(m,\, d,\, t)$ within the k^{th} window, respectively. T is the total number of time-points, K is the total number of nonoverlapping windows and $\omega = T/K$ is the window length. Note that the variance layer works on the outputs x_{conv} of the three convolution layers in parallel.

It is obvious that a high degree of feature reduction is achieved by reducing the number of features to $m \times D \times K$ through the variance layer. In this work, we set the value of K to be 10 for the output x_{var} from previous two variance layers. Particularly, considering the shape of x_{var} from time-frequency convolution followed by the variance layer differs from previous two outputs, the value of K is thus set to be 1. An average pooling layer with size of $(D,\, 1)$ is then applied to reduce its D dimension to 1, which is the same as previous two outputs. Finally, all feature maps are flattened and concatenated into 1D feature vector for the last classifier.

3) Classifier: The classifier includes one FC layer and one softmax layer, which is designed to generate the final decoding results. The 1D feature vector extracted by the TSF-Conv block is fed to the FC layer for classification. The FC layer weights are regularized by using a maximum norm constraint of 0.5, i.e., $\|w\|^2 < 0.5$ (Weight-normalization).

C. Loss Function

The CE loss is adopted to minimize the classification error between network predictions and the ground truth. Moreover, the center loss is used to improve the discriminative power of the deeply learned features. The objective functions of the CE loss and center loss are formulated as:

$$L_{CE} = -\frac{1}{N_b} \sum_{i=1}^{N_b} y_i \, log \, \hat{y}_i \tag{3}$$

$$L_{Center} = \frac{1}{2} \sum_{i=1}^{N_b} \| f_i - c_{y_i} \|_2^2 \tag{4}$$

where y_i is the ground-truth label of the i^{th} training sample, $\hat{y}_i$ is the predicted label of the i^{th} training sample and N_b is the number of samples in a training batch. $f_i \in \mathbb{R}^d$ denotes the feature vector extracted from i^{th} training sample by the network and $c_{y_i} \in \mathbb{R}^d$ denotes the feature center of the class that the sample i belongs to.

As introduced in [22], we update the feature centers c_j in each training iteration as:

$$c_j \leftarrow c_j - \alpha \cdot \Delta c_j \tag{5}$$

$$\Delta c_j = \frac{\sum_{i=1}^{N_b} \delta(i,\, j) \cdot (c_j - f_i)}{1 + \sum_{i=1}^{N_b} \delta(i,\, j)} \tag{6}$$

$$\delta(i,\, j) = \begin{cases} 0, & y_i \neq j \\ 1, & y_i = j \end{cases} \tag{7}$$

where Δc_j is the average distance between the j^{th} class samples and the center vector of the j^{th} class. α denotes the learning rate for center loss, and the value of α is restricted in [0, 1]. The joint supervision of the CE loss and the center loss is advantageous in that minimizing the intra-class variations while keeping the features of different classes separable. Consequently, we obtain the following loss L_{total} to train the network for discriminative feature learning:

$$L_{total} = L_{CE} + \lambda L_{Center} \tag{8}$$

where λ is the trade-off scalar to balance the two loss functions.

In this study, the network supervised by the center loss is optimized by standard SGD [27]. The value of α and λ are set to 0.01 and 0.001, respectively. The influence of different values of α and λ on the performance of model is discussed in Section IV.

III. Experiments and Results

A. Data Description

1) BCI Competition IV 2a Dataset (Dataset I): The dataset [28] contains EEG data from 9 subjects performing four different MI tasks including left hand, right hand, feet and tongue. The signals were recorded from 22 Ag/AgCl electrodes at a sample rate of 250 Hz. Each subject has two sessions and each session has 288 trials, with an average of 72 trials for each class. In this paper, the first session is used for training, and the second session is used for test. The time segment of each trial is restricted between 2s and 6s, which results in 1000 sample points for each trial.

2) BCI Competition IV 2b Dataset (Dataset II): The dataset [29] consists of EEG data from 9 subjects. A total number of 2 MI classes are included: MI of the left hand, right hand. The signals were recorded from 3 electrodes placed at positions C3, Cz, and C4 with the sampling frequency of 250 Hz. For each subject, there are 5 sessions. In this paper, the first three sessions are used for training, and the rest is used for test. There are about 400 trials and 320 trials in the training and test sets respectively. The time segment of each trial is restricted in [3s, 7s], which results in 1000 sample points for each trial.

3) OpenBMI Dataset (Dataset III): The dataset [30] contains EEG signals of 62 channels for 2-class MI tasks recorded from 54 healthy subjects. Following [17], 20 channels in the motor region (FC-5/3/1/2/4/6, C-5/3/1/z/2/4/6, and CP-5/3/1/z/2/4/6) are selected in our experiments.

B. Methods Evaluated

An overview of the benchmark methods is described as follows:

1) FBCSP: FBCSP [8] is a widely used baseline method to decode oscillatory EEG data. This method is based on the combination of bandpass filtering and the CSP algorithm. Note that FBCSP was the best performing method for Dataset I and also won BCI competition IV [28].

2) Deep ConvNet: Deep ConvNet [12] consists of four convolutional layers, with a unique first convolutional layer for spatio-temporal information, followed by three standard convolution-max-pooling blocks and a dense softmax classification layer.

3) EEGNet: EEGNet [13] is a compact CNN for EEG-based extraction of spatial features, and it includes one convolutional layer, one DepthwiseConv2D layer and one SeparableConv2D layer.

4) FBCNet: FBCNet [17] adopts depth-wise convolution to extract spectral-spatial features from a multi-view EEG representation, followed by a variance layer for feature extraction.

5) FBMSNet: FBMSNet [18] is an efficient and lightweight multiscale feature extraction CNN architecture, which extracts multiscale temporal features and spatial features for MI classification.

C. Experimental Setups

1) Experiment Protocols: According to the competition guideline [29], we apply hold-out analysis to evaluate the performance of the TSFCNet, which means that the model is trained and tested completely in different sessions. The specific division manner has been outlined in preceding chapter *Data Description*. This analysis provides information about the capability of the model in extracting highly generalizable discriminative features and tackling the nonstationary phenomenon between two sessions. For fairness, the hold-out analysis is applied for all comparison methods.

2) Training Procedure: As proposed in [12], the training data are further split into a training set and a validation set. During the training process, only the training set is used for training with the early stopping criteria whereby the first phase of the training stops when the validation accuracy does not improve for 200 consecutive epochs. In the second training phase, the training continues on the complete training data, starting from the network parameters that led to the best accuracies on the validation set so far. The training ends when the validation loss drops below the loss of the training set at the end of the first training phase. In this work, the maximum number of training epochs is limited to 1500 and 600 for the two-phase training respectively.

The proposed TSFCNet is implemented with PyTorch 1.12.1 on the NVIDIA GeForce RTX 3090 platform. In addition, the Adam optimizer [31] is employed to optimize the proposed network, and the optimizer parameters β_1 and β_2 are set to 0.9 and 0.999, respectively. The batch size and learning rate of the neural network are set to 32 and 0.001. The center vectors of the center loss are initialized by random Gaussian distribution with a mean of 0 and variance of 1. During the training process, 10% of the training data is set aside as a validation set, and the data in test set would not be used in any of the training phases.

3) Performance Metrics: In the experiments, the classification accuracy (ACC) and the Cohen's kappa coefficient (Kappa) are used as two metrics for performance evaluation. The mathematical formula of Cohen's kappa coefficient is defined as follows:

$$K = \frac{ACC - P_e}{1 - P_e} \tag{9}$$

$$P_e = \frac{\sum_{i=1}^{M} n_{:i} n_{i:}}{N^2} \tag{10}$$

where P_e denotes the hypothetical probability of chance agreement. $n_{:i}$ and $n_{i:}$ are the sum of the i^{th} column and the i^{th} row of the confusion matrix respectively. M is the class number and N is the sum of all entries in the confusion matrix. Wilcoxon signed-rank test is employed to analyze the statistical significance.

D. Performance Comparison

Table I and Table II depict the complete decoding results on both datasets by using the proposed TSFCNet and the other baseline methods. As observed from Table I, the proposed TSFCNet surpasses baseline methods in terms of the average classification accuracy on Dataset I. In particular, the proposed TSFCNet reaches an average accuracy of 82.72%, displaying improvements of 14.97%, 10.04%, 9.26%, 6.56%, and 3.48% over FBCSP ($p < 0.01$), Deep ConvNet ($p < 0.01$), EEGNet ($p < 0.01$), FBCNet ($p < 0.01$) and FBMSNet ($p < 0.05$), respectively. Furthermore, our method achieves higher accuracy on most of subjects except A03, and yields an average kappa value of 0.7695, which is the best among all the methods.

Moreover, Table II illustrates that the proposed TSFCNet outperforms all the SOTA methods in terms of average classification accuracy and kappa value on Dataset II, achieving an accuracy of 86.39% and a kappa value of 0.7324. Additionally, the TSFCNet also shows significant improvement ($p < 0.05$) on the accuracy compared to most of the baseline models. Furthermore, Table I and Table II illustrate that the proposed

TABLE I
CLASSIFICATION PERFORMANCE (%), STD, KAPPA AND P-VALUE ON DATASET I USING OF THE TSFCNET AND THE COMPARED METHODS

Dataset	Methods	A01	A02	A03	A04	A05	A06	A07	A08	A09	Average	Std	Kappa	p-value
I	FBCSP [8]	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75	13.73	0.5700	0.0039
	ConvNet [12]*	78.43	47.53	87.11	65.58	74.61	57.25	78.08	82.94	82.60	72.68	13.22	0.6224	0.0039
	EEGNet [13]*	81.06	59.53	91.13	59.19	70.99	56.76	72.03	83.49	86.96	73.46	12.93	0.6301	0.0039
	FBCNet [17]*	84.38	60.42	93.06	74.65	71.53	53.13	83.33	81.60	83.33	76.16	12.69	0.6821	0.0039
	FBMSNet [18]*	85.42	56.94	**96.18**	81.60	73.26	66.67	84.72	83.68	84.72	79.24	11.73	0.7233	0.0273
	TSFCNet (ours)	**90.28**	**62.50**	93.40	**83.33**	**75.35**	**68.06**	**95.49**	**88.19**	**87.85**	**82.72**	**11.56**	**0.7695**	-

* Reproduced
The bold font highlights the best results among the different methods.

TABLE II
CLASSIFICATION PERFORMANCE (%), STD, KAPPA AND P-VALUE ON DATASET II USING OF THE TSFCNET AND THE COMPARED METHODS

Dataset	Methods	B01	B02	B03	B04	B05	B06	B07	B08	B09	Average	Std	Kappa	p-value
II	FBCSP [8]	70.00	60.36	60.94	**97.50**	93.12	80.63	78.13	92.50	86.88	80.01	13.85	0.6000	0.0279
	ConvNet [12]*	71.56	61.43	74.38	95.63	89.38	81.88	**90.31**	92.50	85.31	82.49	11.28	0.6560	0.0207
	EEGNet [13]*	72.50	58.57	83.13	96.25	86.25	77.81	85.94	**93.13**	79.38	81.44	11.31	0.6295	0.0117
	FBCNet [17]*	75.00	50.36	67.19	96.56	92.81	83.75	77.19	87.81	82.50	79.24	14.10	0.5930	0.0117
	FBMSNet [18]*	72.81	53.21	80.31	97.19	**95.00**	85.63	84.38	90.00	88.75	83.03	13.39	0.6690	0.0273
	TSFCNet (ours)	**76.25**	**70.00**	**83.75**	**97.50**	92.81	**86.56**	88.44	92.50	**89.69**	**86.39**	**8.63**	**0.7324**	-

* Reproduced
The bold font highlights the best results among the different methods.

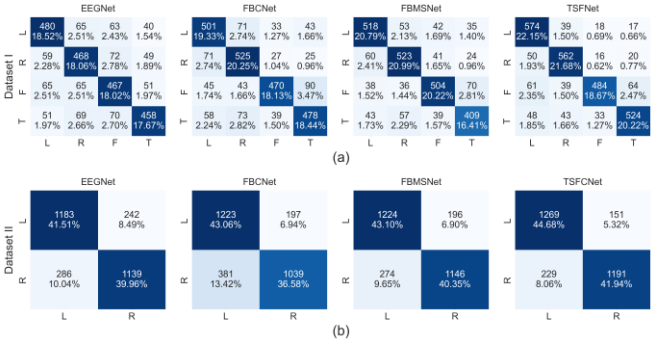

Fig. 2. Confusion matrices on two datasets, where each column represents the actual values and each row depicts the predicted values of the model. L, R, F, and T refer to MI of left hand, right hand, feet, and tongue, respectively. (a) Dataset I. (b) Dataset II.

TABLE III
CLASSIFICATION PERFORMANCE (%), STD AND P-VALUE ON DATASET III USING OF THE TSFCNET AND THE COMPARED METHODS

Methods	ACC (%)	Std	p-value
ConvNet [12]	60.77	11.42	1.7e-08
EEGNet [13]	63.63	**11.08**	1.5e-05
FBCNet [17]	67.19	14.38	0.0002
FBMSNet [18]*	70.20	15.09	0.1013
TSFCNet (ours)	**71.63**	15.58	-

* Reproduced
The bold font highlights the best results among the different methods.

TABLE IV
CLASSIFICATION PERFORMANCE (%), STD, NUMBER OF PARAMETERS AND COMPUTING TIME ON DATASET I USING OF THE TSFCNET AND THE RECENT COMPARED METHODS

Methods	ACC (%)	Std	Params (k)	Computing Time
FBCNet [17]*	76.16	12.69	**11.8**	**0.28**
FBMSNet [18]*	79.24	11.73	30.7	0.49
EEGNeX [19]*	75.15	16.20	62.1	0.29
ATCNet [20]*	80.16	**9.55**	113.7	4.37
Conformer [32]*	76.81	11.55	789.8	0.75
TSFCNet (ours)	**82.72**	11.56	43.1	0.43

* Reproduced
Computing Time: seconds per epoch recorded on a single RTX 3090.
The bold font highlights the best results among the different methods.

TSFCNet has the lowest standard deviation values, which are 11.56 and 8.68, respectively. A graphical representation of the confusion matrix based on Table I and Table II is shown in Fig. 2. In addition to Dataset I and Dataset II, Dataset III is also used for a fair comparison with FBCNet and FBMSNet. Table III demonstrates that the TSFCNet maintains a classification accuracy greater than the baselines and shows significant improvement compared to most of them.

Particularly, we also conduct comparative experiments on the classification accuracy, model complexity and computing efficiency with most recent studies [17] [18] [19] [20] [32]. As Table IV shows, the TSFCNet achieves better classification accuracy and relatively low standard. Notably, the TSFCNet exhibits competitive computing efficiency, with an average time of 0.43 seconds per epoch. Furthermore, the TSFCNet

TABLE V
CLASSIFICATION PERFORMANCE (%), STD, KAPPA AND P-VALUE ON DATASET I UNDER10-FOLD CROSS VALIDATION SCENARIO

Methods	ACC (%)	Std	Kappa	p-value
3D-CNN [33]	75.02	7.34	0.6440	0.0039
Discriminative Feature [34]	81.85	10.77	0.7580	0.0039
TS-SEFFNet [35]	84.49	-	0.7940	-
FBMSNet [18]*	87.04	8.55	0.8272	0.0039
EEGNeX [19]*	88.12	7.24	0.8413	0.0039
TSFCNet (ours)	**89.49**	**7.16**	**0.8609**	-

* Reproduced

The bold font highlights the best results among the different methods.

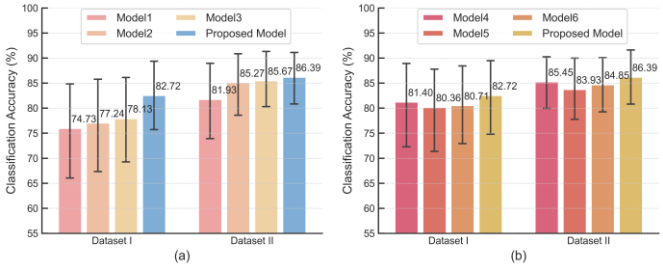

Fig. 3. The classification accuracy comparison of the ablation study on two Datasets. (a) the first ablation study. (b) the second ablation study.

strikes a balance between model complexity and performance. This analysis demonstrates the efficacy of the TSFCNet, as it achieves superior classification accuracy and efficiency compared to the SOTA models.

Additionally, several recent studies [19] [33] [34] [35] have presented outcomes for Dataset I through the utilization of 10-fold cross-validation on merged data (576 trials) for each subject. Hence, to ensure the fair comparison with these deep learning methods, Table V presents a comparative analysis between the proposed TSFCNet and recent studies on Dataset I by evaluating on merged data (576 trials) for each subject. As shown in Table IV, in 10-fold cross-validation scenario, the accuracy, standard deviation value and kappa value of the proposed TSFCNet is 89.49%, 7.16 and 0.8609, respectively. The results indicate that the performance of our method is higher than that of the competing methods.

As a result, the experimental results on three classical datasets demonstrate that the proposed TSFCNet achieves impressive performance and promising robustness for MI-EEG decoding.

E. Result of Ablation Experiments

To verify the effectiveness of integrating MixConv-Residual block and center loss into the TSFCNet model, the first ablation study is conducted on Dataset I and Dataset II, as shown in Fig. 3(a). Three models, named Model1, Model2 and Model3 are utilized represent three scenarios as follows:

1) Model1: The model is implemented by removing the MixConv-Residual block from the TSFCNet and trained with CE loss.

2) Model2: The model is implemented by integrating MixConv layer into the Model1 for extracting multiscale temporal information, and also trained with CE loss.

3) Model3: The complete TSFCNet model with MixConv-Residual block is trained with CE loss.

Fig. 3(a) shows the classification accuracies and standard deviation values obtained from the first ablation study. It can be seen that the application of the MixConv layer leads to a substantial 2.51% accuracy improvement for the Model2 on Dataset I, due to the extraction of multiscale temporal information. Additionally, the employment of the residual mechanism results in a 0.89% accuracy improvement for the Model3 over the Model2. The similar result is also observed to on Dataset II, where the use of the MixConv-Residual block yields a more significant improvement of 3.74% in accuracy. By adding the center loss as an auxiliary cost into the Model3, the proposed TSFCNet is able to improve the decoding accuracy in a step further on two datasets. Notably, the proposed TSFCNet can reach 4.59% higher accuracy than the model3 on Dataset I. Furthermore, a decreasing trend in the standard deviation values indicates that the MixConv-Residual block and the center loss strategy could not only improve the performance on EEG classification but also the robustness of the model.

Additionally, for further investigating the importance and contributions of each convolutional operation in the TSF-Conv block, we propose another three simplified models to conduct the second ablation study, which are introduced as follows:

4) Model4: The model is implemented without the spatial convolutional operation in the TSF-Conv block and trained with the center loss.

5) Model5: The model is implemented without the frequency convolutional operation in the TSF-Conv block and trained with the center loss.

6) Model6: The model is implemented without the time-frequency convolutional operation in the TSF-Conv block and trained with the center loss.

Fig. 3(b) shows the classification accuracies and standard deviation values obtained from the second ablation study. In general, the proposed TSFCNet outperforms Model4, Model5 and Model6 on two datasets, which demonstrates that the absence of any convolutional operation in the TSFConv block leads to a decline in classification performance. Particularly, Model4 has the least effect on the classification accuracy, while the performance still lags behind the proposed TSFCNet. Removing the frequency convolutional operation from the TSFConv block leads to a significant decrease of accuracy in the Model5 result, with a reduction of 2.36% and 2.46% on Dataset I and Dataset II, respectively. It indicates that the frequency convolutional operation is crucial in capturing highly discriminative representations. On the other hand, Model6 also leads to a significant decrease of accuracy in the absence of the time-frequency convolutional operation. Furthermore, in the second ablation study, the proposed TSFCNet also achieves the lowest standard deviation. These experimental results highlight that the proposed TSFCNet with the complete TSFConv block can effectively capture the essential spatial, frequency and time-frequency feature representations.

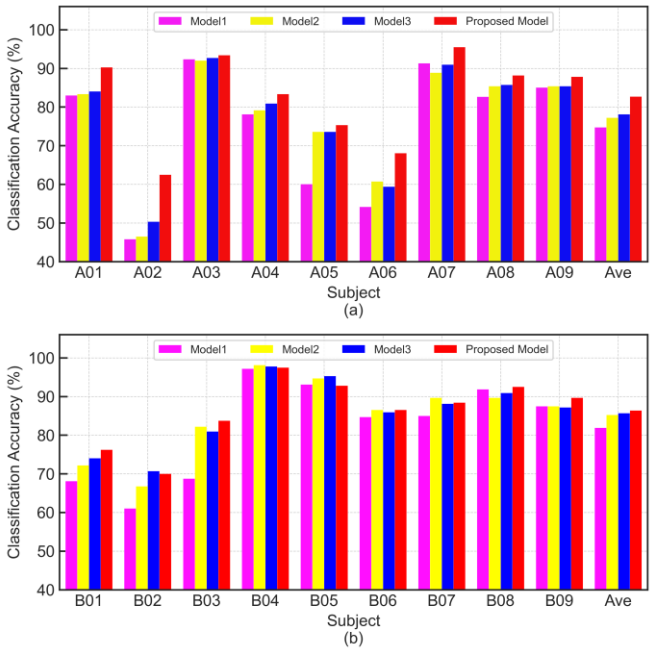

Fig. 4. The accuracy comparison of each subject in the first ablation study on two datasets. (a) Dataset I. (b) Dataset II.

IV. Discussions

A. Efficacy of MixConv-Residual Block and Center Loss

Figure 4 shows the classification accuracy of each subject in the first ablation study, which has verified the effects of the MixConv-Residual block and center loss on model accuracy. The EEG data offers abundant temporal information due to its high temporal resolution. However, as shown in Figure 4, Model1 displays unsatisfactory classification performance due to a lack of temporal feature extraction. To tackle this problem, the MixConv layer for multiscale temporal feature extraction is applied and has improved the classification accuracy of Model2 significantly. Specifically, the proposed method has achieved a notable increase in classification accuracy in subjects A04, A05, A06, and A08 on Dataset I. At the same time, most of subjects achieves an improvement except for B08 on Dataset II. Note that since the EEG data on Dataset II has fewer channels, the TSF-Conv block can only extract fewer features. Therefore, the effect of multiscale temporal feature extraction is more pronounced. Although the MixConv layer may decrease the classification accuracy of a few subjects, the influence is generally small. These analyses demonstrate the importance and effectiveness of temporal feature extraction in EEG data decoding.

Moreover, we incorporate the residual mechanism into the MixConv layer to form the MixConv-Residual block. The residual mechanism has limited effects on the overall classification performance, with significant improvements only observed for subjects A02, A04 on Dataset I and subjects B01, B02 and B05 on Dataset II. The MixConv-Residual block can prevent learning degradation and reduce temporal feature extraction redundancy, as seen in the improved classification accuracy of subjects A03, A07 and B08.

Finally, the proposed TSFCNet outperforms the SOTA classification performance by introducing the center loss as an auxiliary loss function. On Dataset I, all subjects trained with the center loss gain significant improvements in classification accuracy, particularly subjects A01, A02, and A06. The average classification results on Dataset I are higher with center loss training. Additionally, the classification accuracy on Dataset II is also improved. It is notable that the effects of the center loss in the four-class classification are superior to those in the two-class classification. These experimental results indicate that the center loss could make the samples that belong to the same class compact in the feature space, which could significantly improve the MI decoding. Therefore, by employing the novel MixConv-Residual block and the center loss, the proposed TSFCNet is able to obtain more discriminative temporal information at different scales which result in the increase of classification accuracy.

B. Efficacy of TSF-Conv Block

The TSF-Conv block implements three parallel and independent convolutional operations to extract multi-domain high discriminative features in the spatial, frequency and time-frequency domain. It is different from the deep and complicated CNN architectures that generally focus on the deep feature in limited domains (temporal and spatial, or both). The TSF-Conv block is effective, intuitional and simple. Such kind of architecture could avoid the marginal effect and redundant or irrelative information, and hence improve the quality and interpretability of the EEG decoding. As shown in Fig. 3(b), although the spatial convolutional operation contributes to providing spatial information for feature representations, the second ablation study shows that its effect is limited. Inspiringly, feature extraction of frequency domain by convolutional operation directly could significantly improve EEG decoding ability. Recent researches have shown that spatial-temporal convolution on the spectrally filtered EEG data only enhances spatial-temporal features on different frequencies. Compared with FBCNet and FBMSNet, the main improvements of the TSFCNet are introducing residual mechanism and TSF-Conv block. The TSFCNet provides a novel perspective that using frequency and time-frequency features to improve the EEG decoding in new domains. The shallow and effective nature of these three convolutional operations in TSF-Conv block along with fewer parameters reduces the training time, and this multi-domain framework could guide the design of CNN structures for EEG decoding.

Additionally, inspired by [17], the variance layer is used to extract and compress the temporal features obtained from the preceding TSF-Conv layers. Such kind of variance operation along the time domain is suitable for temporal consolidation since the variance of a filtered signal could be consider as the spectral power in the time-series. Consequently, combining with the multiscale temporal features extracted from the MixConv-Residual Block, the TSF-Conv block gains the final SOTA classification performance.

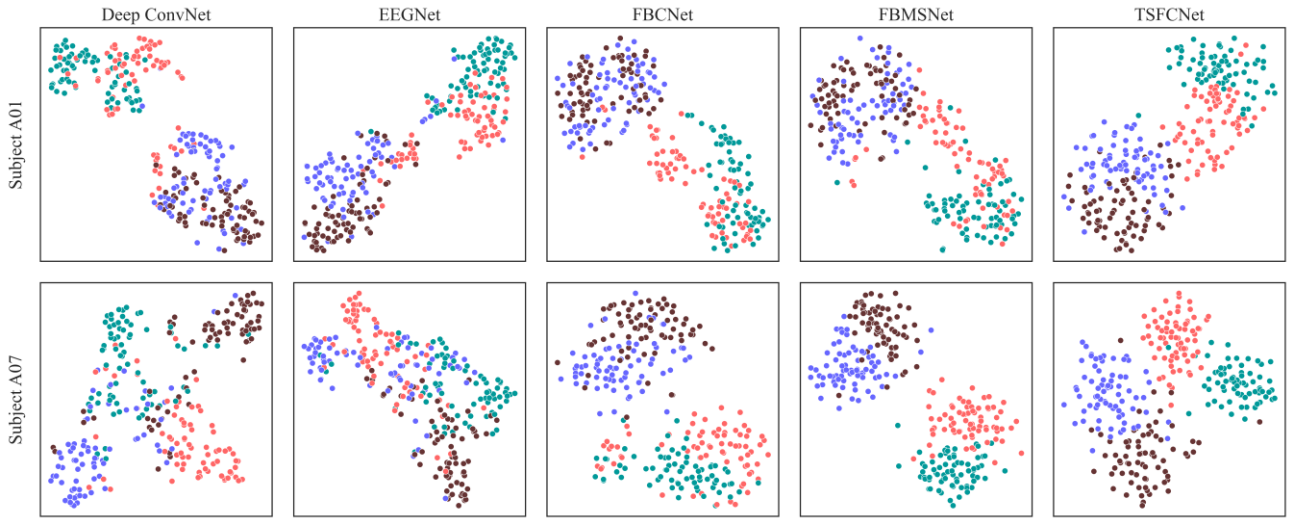

Fig. 5. Comparison of the features from subject A01 and A07 on Dataset I learned by different methods in the 2-D embedding space by t-SNE. Red, green, blue and brown points represent the MI of left hand, right hand, foot and tongue, respectively.

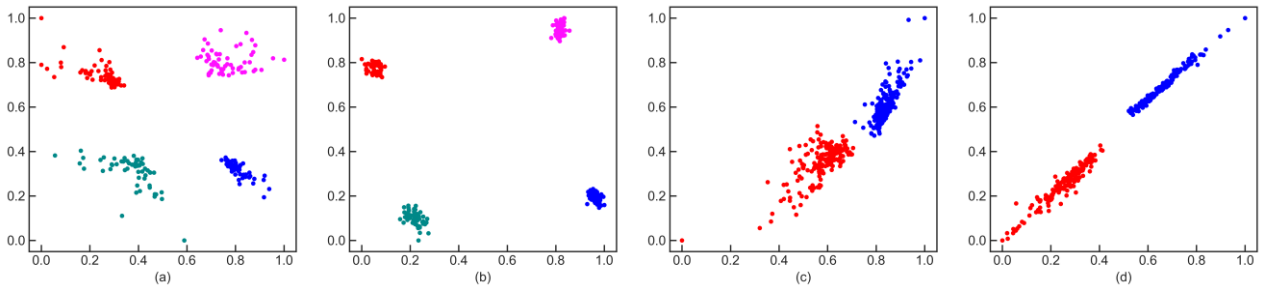

Fig. 6. The distribution of deeply learned features under the supervision of different loss on two datasets. The points with different colors denote features from different classes. (a) the supervision of CE loss on Dataset I. (b) the joint supervision of CE loss and center loss on Dataset I. (c) the supervision of CE loss on Dataset II. (d) the joint supervision of CE loss and center loss on Dataset II.

C. Visualization

To make further discussion on the discriminatory capabilities of the features derived from the proposed TSFCNet, we use t-SNE [36] to produce a two-dimensional embedding of the learned EEG features. The resultant visualization is presented in Fig. 5.

As shown in Fig. 5, the visualizations of features extracted from Deep ConvNet, EEGNet, and FBCNet exhibit a large degree of overlap between different classes, resulting in ambiguity in their classification. In contrast, the proposed TSFCNet demonstrates better performance in capturing distinct features from MI-EEG and achieving minimal overlap. Specifically, the proposed TSFCNet could generate a higher degree of inter-class distance and a lower degree of intra-class distance compared to FBMSNet. This is achieved by the incorporation of multiscale temporal features and multi-domain TSF-Conv features. It enables the efficient discrimination of various types of MI-EEG signals. Consequently, our results demonstrate that the TSFCNet is capable of extracting highly discriminative EEG features, leading to improved decoding performance.

Additionally, the visualization method proposed in [22] is employed to examine the distribution of deeply learned features under the supervision of different losses on two datasets. Specifically, the TSFCNet is modified by reducing the output of the last hidden layer to a 1×2 vector, thereby allowing for direct visualization of the features on a two-dimensional surface.

As shown in Fig. 6(a)(c), we could observe that the deeply learned features are separable under the supervision of the CE loss. However, the deep features are still not sufficiently discriminative due to their significant intra-class variances. Conversely, Fig. 6(b)(d) depicts that the deeply learned features exhibit greater intra-class compact and increases inter-class distance under the joint supervision of CE loss and center loss. It indicates that the center loss method can enhance the discriminative ability of feature vectors, which also can be proved by the first ablation study results.

As shown in Fig. 7, the visualization of learned weights on the EEG topography is achieved by employing Gradient-weighted Class Activation Mapping (Grad-CAM) [37] in the proposed model. The obtained Grad-CAM results demonstrate that the presence of contralateral activation patterns in accordance with the paradigm of motor imagery [38]. The observed activation patterns (red color) in the motor-related areas of the left and right hemispheres are consistent with the imagined right-hand and left-hand movements. Our model builds the explicit connection between these activated regions and the decision-making process. The ability of aligning the

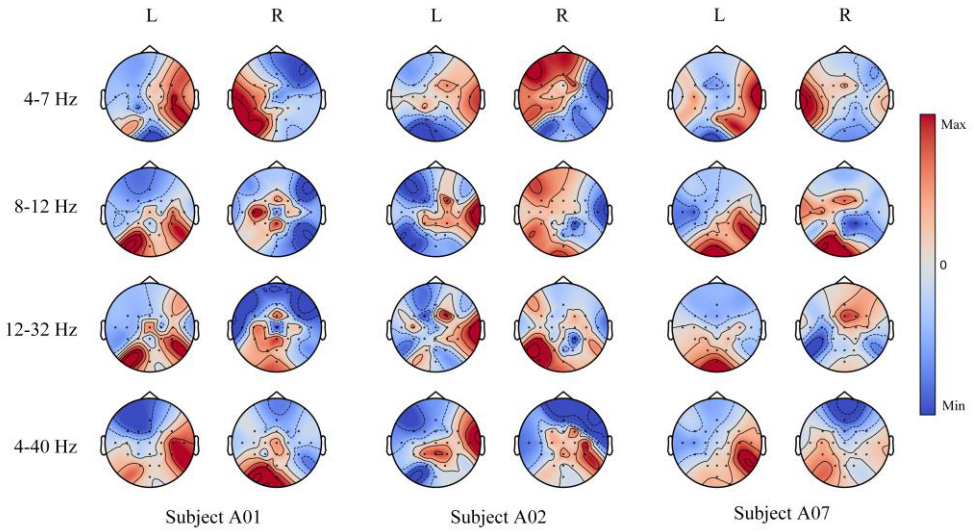

Fig. 7. Gradient-weighted Class Activation Mapping (Grad-CAM) of the proposed TSFCNet on the head EEG topography. Contralateral activation patterns can be clearly observed in the subject A01, A02 and A07. (L: left-hand MI, R: right-hand MI).

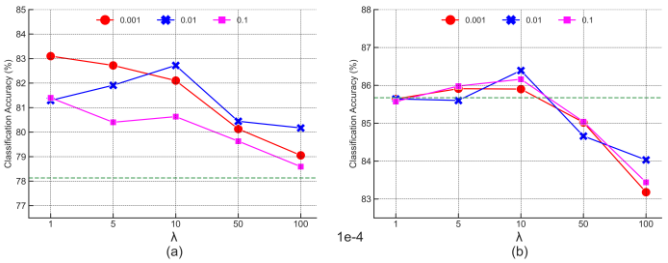

Fig. 8. The classification accuracy of TSFCNet across various settings of α and λ on two datasets, where the green dashed line represents the accuracy by TSFCNet with $\lambda = 0$. (a) Dataset I. (b) Dataset II.

feature distributions learned by the proposed TSFCNet with different MI tasks across various frequency bands can reveal potential associations between body movements and brain activities, making it a valuable tool.

D. Influence of Values of α and λ

The TSFCNet utilizes the joint supervision of the CE loss and the center loss through Eq. (8) for model training. The effective implementation of the center loss requires careful consideration of two key hyperparameters, namely the learning rate of feature centers α and the trade-off value λ between the CE and the center loss. In order to evaluate the influence of the aforementioned hyperparameters, an empirical investigation is conducted to compare the performance of the TSFCNet across various settings of α and λ on both Dataset I and Dataset II.

As shown in Fig. 8(a), the classification accuracy exhibited by the TSFCNet on Dataset I is observed to decrease with the increase of the value of λ, except for the scenario when α equals 0.01 and $\lambda \in [0.0001, 0.001]$, where an increase in decoding accuracy is noted. In Fig. 8(b), it is demonstrated that the increase of the value of λ from 0 to 0.001 leads to an improvement in the classification accuracy of the TSFCNet on Dataset II. Notably, discriminative features are observed for varies values of λ between 0.0001 and 0.001. However, a further escalation in the λ value causes the decline of the MI decoding performance of the TSFCNet. Moreover, the trend of the accuracy curve across varying learning rate is similar, which is in line with the result in [22]. Consequently, in this work, we experimentally set the values of α and λ to 0.01 and 0.001, respectively. Therefore, the satisfactory decoding performances on both Dataset I and Dataset II are obtained.

E. Limitation and Future Work

Despite the proposed TSFCNet achieves qualified and robust decoding results, our present work still has some limitations. First, the proposed TSFCNet employs three independent single layers in multi-domain for feature extraction, which may lead to the neglect of the deep information that could improve the classification performance. Second, although the proposed TSFCNet shows its effectiveness in decoding subject-specific MI-EEG, the generalizability across different subjects requires further investigation. Third, the proposed TSFCNet is an offline neural network that is yet to be validated in online BCI environments.

The deep neural networks have shown remarkable capability in absorbing extensive datasets for generating better feature representations. On the other hand, the effective utilization of cross-subject tasks and data augmentation techniques can provide more available training data for deep neural networks. With the help of the deep neural networks and big data, it is possible for conducting the online BCI experiments. Therefore, in the future work, we will explore the potential of the TSFCNet by developing deep architectures for cross-subject tasks.

V. Conclusion

In this paper, a multi-domain temporal-spatial-frequency convolutional neural network is proposed for MI-EEG decoding. The proposed TSFCNet first extracts multiscale temporal feature from filtered EEG signals via the MixConv-Residual Block. Next, the TSF-Conv block learns discriminative

multi-domain EEG presentations through three parallel and independent convolutional operations. In addition, the proposed TSFCNet provides a novel mechanism that leverages spatial and temporal EEG features combined with frequency and time-frequency characteristics to improve the EEG decoding. It enables powerful feature extraction without complicated network structure. Moreover, we combine the center loss with the CE loss to enhance the discriminative abilities of feature extraction. The results of our experiments, conducted on three public BCI datasets, demonstrate that the performance of the TSFCNet is better than that of the SOTA methods. The code of the TSFCNet can be accessed freely. In conclusion, the experimental results demonstrate the proposed method to be efficient and robust in decoding MI-EEG signals and prove it as a powerful tool for MI-EEG based BCIs.

REFERENCES

[1] D. Wen et al., "Combining brain–computer interface and virtual reality for rehabilitation in neurological diseases: A narrative review," *Ann. Phys. Rehabil. Med.*, vol. 64, no. 1, Jan. 2021, Art. no. 101404.

[2] R. Xu et al., "A closed-loop brain–computer interface triggering an active ankle–foot orthosis for inducing cortical neural plasticity," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 7, pp. 2092–2101, Jul. 2014.

[3] R. Chai, S. H. Ling, G. P. Hunter, Y. Tran, and H. T. Nguyen, "Brain–computer interface classifier for wheelchair commands using neural network with fuzzy particle swarm optimization," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 5, pp. 1614–1624, Sep. 2014.

[4] L. He, D. Hu, M. Wan, Y. Wen, K. M. von Deneen, and M. Zhou, "Common Bayesian network for classification of EEG-based multiclass motor imagery BCI," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 46, no. 6, pp. 843–854, Jun. 2016.

[5] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain–computer communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001.

[6] 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, Apr. 2018, Art. no. 031005.

[7] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[8] K. Keng Ang, Z. Yang Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain–computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intelligence*, Jun. 2008, pp. 2390–2397.

[9] K. P. Thomas, C. Guan, C. T. Lau, A. P. Vinod, and K. K. Ang, "A new discriminative common spatial pattern method for motor imagery brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 11, pp. 2730–2733, Nov. 2009.

[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] 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.

[13] 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.

[14] Y. Li, X.-R. Zhang, B. Zhang, M.-Y. Lei, W.-G. Cui, and Y.-Z. Guo, "A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1170–1180, Jun. 2019.

[15] 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.

[16] C. Zhang, Y.-K. Kim, and A. Eskandarian, "EEG-inception: An accurate and robust end-to-end neural network for EEG-based motor imagery classification," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 046014.

[17] R. Mane et al., "FBCNet: A multi-view convolutional neural network for brain–computer interface," 2021, *arXiv:2104.01233*.

[18] K. Liu, M. Yang, Z. Yu, G. Wang, and W. Wu, "FBMSNet: A filter-bank multi-scale convolutional neural network for EEG-based motor imagery decoding," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 2, pp. 436–445, Feb. 2023.

[19] X. Chen, X. Teng, H. Chen, Y. Pan, and P. Geyer, "Toward reliable signals decoding for electroencephalogram: A benchmark study to EEGNeX," 2022, *arXiv:2207.12369*.

[20] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for EEG-based motor imagery classification," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 2249–2258, Feb. 2023.

[21] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Dynamic convolution with multilevel attention for EEG-based motor imagery decoding," *IEEE Internet Things J.*, Feb. 2023, doi: 10.1109/JIOT.2023.3281911.

[22] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 499–515.

[23] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[24] 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.

[25] M. Tan and Q. V. Le, "MixConv: Mixed depthwise convolutional kernels," 2019, *arXiv:1907.09595*.

[26] 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.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[28] M. Tangermann et al., "Review of the BCI competition IV," *Frontiers Neurosci.*, vol. 6, p. 55, Jul. 2012.

[29] 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, p. 39, Mar. 2012.

[30] 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.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[32] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 31, pp. 710–719, Dec. 2023.

[33] 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.

[34] L. Yang, Y. Song, K. Ma, and L. Xie, "Motor imagery EEG decoding method based on a discriminative feature learning strategy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 368–379, Jan. 2021.

[35] Y. Li, L. Guo, Y. Liu, J. Liu, and F. Meng, "A temporal-spectral-based squeeze-and-excitation feature fusion network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1534–1545, Jul. 2021.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[37] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2017, pp. 618–626.

[38] 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.