A Temporal-Spectral-Based Squeeze-and-Excitation Feature Fusion Network for Motor Imagery EEG Decoding

Yang Li, Lianghui Guo, Yu Liu, Jingyu Liu, and Fangang Meng

***Abstract*— Motor imagery (MI) electroencephalography (EEG) decoding plays an important role in brain-computer interface (BCI), which enables motor-disabled patients to communicate with the outside world via external devices. Recent deep learning methods, which fail to fully explore both deep-temporal characterizations in EEGs itself and multi-spectral information in different rhythms, generally ignore the temporal or spectral dependencies in MI-EEG. Also, the lack of effective feature fusion probably leads to redundant or irrelative information and thus fails to achieve the most discriminative features, resulting in the limited MI-EEG decoding performance. To address these issues, in this paper, a MI-EEG decoding framework is proposed, which uses a novel temporal-spectral-based squeeze-and-excitation feature fusion network (TS-SEFFNet). First, the deep-temporal convolution block (DT-Conv block) implements convolutions in a cascade architecture, which extracts high-dimension temporal representations from raw EEG signals. Second, the multi-spectral convolution block (MS-Conv block) is then conducted in parallel using multi-level wavelet convolutions to capture discriminative spectral features from corresponding clinical subbands. Finally, the proposed squeeze-and-excitation feature fusion block (SE-Feature-Fusion block) maps the deep-temporal and multi-spectral features into comprehensive fused feature maps, which highlights channel-wise feature responses by constructing interdependencies among different domain features. Competitive experimental results on two public datasets demonstrate that our method is able to achieve promising decoding performance compared with the state-of-the-art methods.**

Manuscript received April 9, 2021; revised June 25, 2021; accepted July 19, 2021. Date of publication July 26, 2021; date of current version August 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant U1809209 and Grant 61671042 and in part by the Beijing Natural Science Foundation under Grant L182015. *(Corresponding authors: Lianghui Guo; Yu Liu.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Yang Li is with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Department of Automation Sciences and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: liyang@buaa.edu.cn).

Lianghui Guo, Yu Liu, and Jingyu Liu are with the Department of Automation Science and Electrical Engineering, Beihang University, Beijing 100083, China (e-mail: guolianghui_buaa@163.com; sy1803113@buaa.edu.cn; liujingyu@buaa.edu.cn).

Fangang Meng is with the Beijing Key Laboratory of Neuroelectrical Stimulation Research and Treatment, Beijing Institute of Neurosurgery, Capital Medical University, Beijing 100069, China (e-mail: mengfangang@126.com).

Data is available on-line at https://github.com/LianghuiGuo/TS-SEFFNet.

Digital Object Identifier 10.1109/TNSRE.2021.3099908

***Index Terms*— EEG, motor imagery, deep-temporal convolution, multi-spectral convolution, squeeze-and-excitation feature fusion.**

I. Introduction

THE brain computer interface (BCI) creatively provides new ways for communication between human and computers by analyzing the electric signals generated by brain and translating them into real commands [1], which helps to control the external devices. With fast development of human-computer interaction technology, MI-EEG signals are widely used in BCI researches, which study the triggered neural activities in brain areas relevant to imagery body movements [2]. If these imagery-movement based neural activities are decoded correctly, people with severe motor diseases can control external devices via the decoded MI-EEG signals [3]. Therefore, the MI-EEG based pattern recognition and correct decoding are important in these BCI systems. However, it is difficult to realize effective MI-EEG decoding due to the low signal to noise ratio (SNR), non-stationarity in signals and individual differences of subjects [4].

In order to realize MI-EEG decoding, many machine learning algorithms have been proposed [5]–[9]. For example, the Common Spatial Pattern (CSP) was a powerful algorithm in discriminating MI-EEG signals [6], which constructed spatial filters and extracted time-frequency features effectively. Ang *et al.* [7] utilized filter bank CSP (FBSCP) to extract the optimal spatial features from a group of bandpass filters. Saha *et al.* further [10] used CSP with Joint Approximate Diagonalization (JAD) and applied wavelet decomposition as the feature extraction method, which generated the subband energy and entropy of EEG. However, these methods above highly focused on the energy features of EEG, which failed to obtain features with high discrimination from raw EEG signals subject-dependently, and thus limited the decoding performance of MI-EEG [9].

Deep learning algorithms tackled the above problems to some extent by exploiting MI-EEG patterns in a data driven manner. For instance, Chen *et al.* [11] developed a decoding framework including a filter bank spatial filtering and a designed convolutional neural network (CNN). Zhao *et al.* [12] also built a multi-branch 3D CNN framework by transforming EEG into series of 2D array, which focused on the spatial distribution of electrode signals. Zhang *et al.* [13] further designed a hybrid network for spatial and temporal feature

extraction. However, the conventional convolution operation behaves essentially similar to the low-pass filter, which leads to the loss of component in high frequency bands [14]. These deficiencies indicated that incomplete temporal or spatial analysis in end-to-end CNN framework perhaps undermined the severe non-stationarity of EEG. In summary, recent deep learning methods were limited in MI-EEG decoding since spectral, temporal and spatial information are seldom considered simultaneously and discriminative features are not extracted effectively.

In order to address these weaknesses, some recent studies started to explore multi-domain information in EEG recognition. Sakhavi *et al.* [15] utilized a FBCSP to generate temporal representations of EEG which were then fed into a CNN for the EEG classification. In addition, the spiking neural network was used [16] and combined with OVR (One-Vs-Rest) FBCSP which extracted temporal-frequency features from multiple MI-EEG. Li *et al.* [14] further investigated multi-domain representation of EEG and performed spectral and temporal analysis together with a designed CNN. However, some limitations still exist in these methods. First, most widely used methods decoded MI-EEG based on shallow feature extraction, and the absence of deeper information leaded to inaccurate decoding results. Second, these methods tended to investigate temporal or spectral features separately and combined the captured features directly, which probably leaded to information redundancy since they neglected the importance of effective feature fusion. Consequently, it is far from enough to blindly combine parallel algorithms for discriminative feature extraction. Recently, the attention unit named squeeze-and-excitation (SE) was developed to emphasize the informative channel-wise feature [17]. Inspired by its advantages, we aim to adopt the highly recognition performance of the SE to boost the most discriminative temporal and spectral features and further implement feature fusion effectively.

To address the issues above, in this paper, we develop a novel end-to-end MI-EEG decoding framework, named a temporal-spectral-based squeeze-and-excitation feature fusion network (TS-SEFFNet), which involves five subblocks as follows. First, raw MI-EEG signals are embedded into preliminary representations via the spatio-temporal block, which extracts coarse temporal representation and spatial dependencies simultaneously. Second, the deep-temporal convolution block (DT-Conv block) further employs several temporal conv units to capture crucial dynamic temporal features from a higher level. Meanwhile, the multi-spectral convolution block (MS-Conv block) aims to obtain multi-spectral EEG representations corresponding to certain clinical frequency subbands. EEG signals are further mapped into deep-temporal and multi-spectral representations parallelly by the previous feature extracting blocks. Moreover, in order to effectively fuse the temporal-spectral features, the squeeze-and-excitation feature fusion block (SE-Feature-Fusion block) is designed to apply higher weights on more discriminative feature maps when implementing the feature fusion, which alleviates the problem of feature redundancy compared to the direct fusion. Finally, the fused features are fed into the classification block. The proposed TS-SEFFNet is evaluated on two public datasets and gains better performance compared with the state-of-the-art algorithms, demonstrating its efficacy in MI-EEG decoding.

Main contributions of this paper are summarized as follows:

1) We propose a novel TS-SEFFNet for MI-EEG decoding, which integrates both deep-temporal and multi-spectral feature extraction into the deep learning model simultaneously. Particularly, the proposed TS-SEFFNet captures features of MI-EEG more sensitively and accurately than the widely used spatio-temporal-based model.

2) We design a DT-Conv block to extract high-dimensional information from MI-EEG, which generates important dynamic temporal features of EEG signals and avoids the loss of high-level representation in shallow networks.

3) We introduce a MS-Conv block to implement powerful spectral feature extraction and supplement the decoding model with discriminative multi-spectral information, which effectively solves the problem of insufficient feature extraction in single spectral networks.

4) An attention-based fusion method named SE-Feature-Fusion is adopted to alleviate the problem of information redundancy caused by direct fusion, which efficiently extracts the most discriminative features while suppressing the less informative features, and thus boosts the decoding performance.

II. Methodology

This section describes the notations and definitions used in this paper. First, we give a detailed interpretation of the proposed TS-SEFFNet and describe each subblock of the whole net, including spatio-temporal block, DT-Conv block, MS-Conv block, SE-Feature-Fusion block and classification block, respectively. Then, we show the overall scheme of the MI-EEG decoding process. Finally, we summarize the proposed TS-SEFFNet.

A. Notations and Definitions

The raw EEG signals are defined as $E = \{(X_i, y_i)|i = 1, 2, \ldots, N\}$, where $X_i \in R^{C\times K}$ is a two-dimension array representing the i-th EEG trial with C channels and K samples. N is the total number of EEG signal trials. y_i is the corresponding label of X_i and takes its value from set L which contains M classes in a motor imagery task. For example, a four-type motor imagery dataset contains its corresponding label set: $L = \{l_1 = \text{“}left\text{”}, l_2 = \text{“}right\text{”}, l_3 = \text{“}feet\text{”}, l_4 = \text{“}rest\text{”}\}$. The shape of the feature map in the model is defined as $(m\ @\ c \times t)$, describing number, width and length of the feature map respectively. The size of each convolutional filter kernel is denoted as $c \times t$, where c is the channel dimension and t is the time dimension.

B. Temporal-Spectral-Based Squeeze-and-Excitation Feature Fusion Network

In this section, we detailly demonstrate our proposed TS-SEFFNet, which is showed in Fig. 1.

1) The Design of the Spatio-Temporal Block: EEG signals contain abundant temporal, spatial and spectral features which are difficult to define manually [18]. Therefore, as the first part of the proposed decoding framework, the spatio-temporal

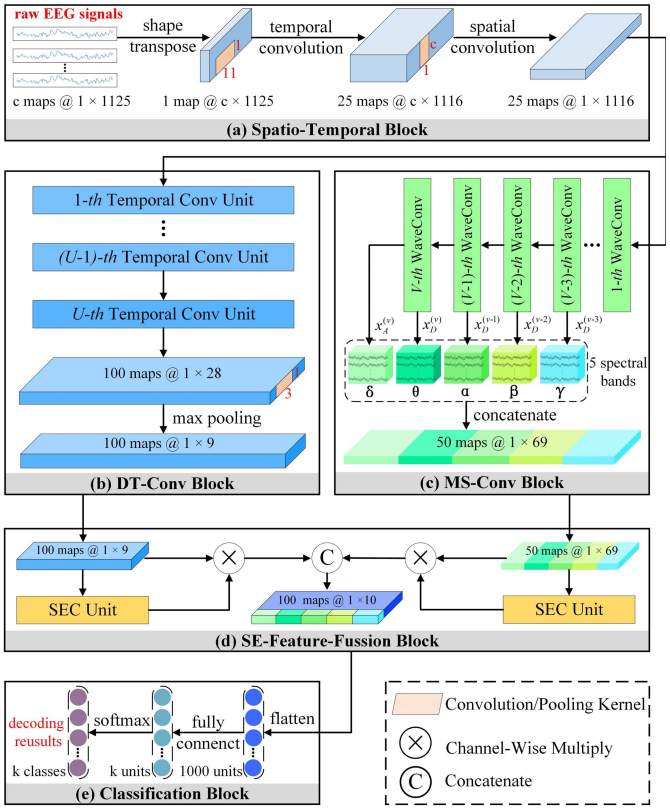

Fig. 1. An illustration of the proposed TS-SEFFNet architecture.

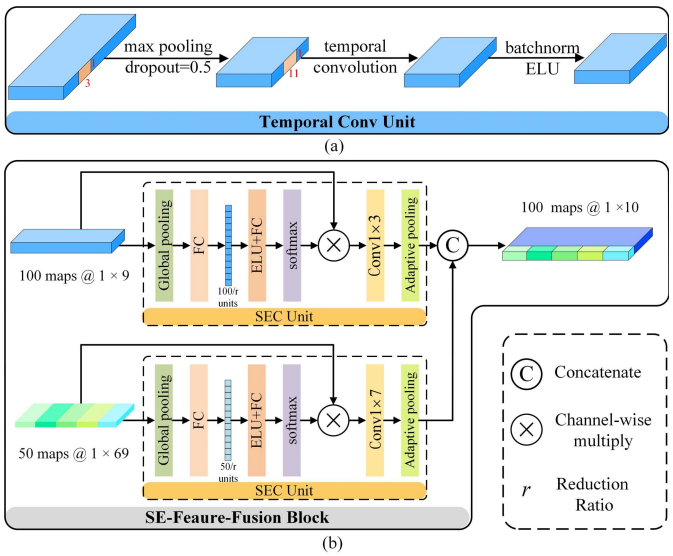

Fig. 2. The architecture of the proposed (a) Temporal Conv Unit, (b) SE-Feature-Fusion Block.

block applies CNN to directly extract preliminary features from EEG signals. Fig. 1(a) shows the structure of the spatio-temporal block, which includes the shape transforming layer, the temporal convolution layer and the spatial convolution layer. Raw EEG signals are firstly transformed from the original temporal representation into 2D maps. Additionally, in order to extract spatial features and temporal dependencies from raw EEG signals, the spatio-temporal block firstly performs a convolution over time with the kernel size 1×11 [19], and then the second convolution implements a spatial filtering with the size of $c \times 1$ over all channels. These two layers implicitly transform EEG signals into a combination of temporal and spatial representation. Exponential linear unit (ELU) [19] and batch-normalization are further adopted after the spatial convolution layer. As a result, the spatio-temporal block generates a set of low-level EEG representations, which are then fed into both MS-Conv block and DT-Conv block for further feature extraction.

2) The Design of the DT-Conv Block: Since the above spatio-temporal features obtained are relatively coarse and not informative enough, deeper feature extraction method is highly necessary. In this block, in order to further explore deeper temporal features, a DT-Conv block is designed by using successive convolution units among time dimension.

Fig. 1(b) is the layout of the DT-Conv block, which includes U designed units named Temporal Conv unit, and the structure of each unit is given in Fig. 2(a). Each unit starts from a max-ing pooling layer with the size of 1×3, generating a coarser representation of EEG features by down-sampling [20]. Then a temporal convolution with a kernel size 1×11 is employed, which aims at producing deeper scale information. Additionally, dropout and batch-normalization operations are adopted in each unit which alleviate the overfitting problem caused by the inadequacy of EEG data [21]. The DT-Conv block arranges all the units in sequence, and deeper EEG representations are then extracted from the elementary shallow spatio-temporal features. The number of units U and the length of kernel size are tuned by the classification performance, which are demonstrated in Section-III in detail.

3) The Design of the MS-Conv Block: The efficient extraction of EEG spectral component is important in MI-based BCI, which is particularly challenging due to the non-stationarity of EEG [22]. Therefore, in order to integrate spectral analysis into an end-to-end model without increasing computational cost, a MS-Conv block is proposed by using a series of wavelet convolutions to obtain multi-spectral representations corresponding to five clinical bands and further concatenate them into multi-spectral features.

Concretely, Fig. 1(c) shows the layout of the MS-Conv block. In order to integrate the spectral feature extraction into the proposed TS-SEFFNet, we design a convolution operator named wavelet convolution (WaveConv), which implements the wavelet decomposition on EEG representation via convolution layer [14]. Daubechies order-4 (Db4) wavelet has good orthogonality property and efficient filter implementation [23], which involves no learnable parameters in a WaveConv. Previous study also reported that Db4 wavelet is useful for the spectral feature extraction due to its high correlation coefficients with brain signals [14], and thus Db4 is chosen in this paper. After a series of WaveConv layers, the EEG representations are decomposed into coefficients corresponding to five frequency subbands that satisfy the clinical interests [24], [25]: 0-4Hz (δ rhythm), 4-8Hz (θ rhythm), 8-12Hz (α rhythm), 13-30Hz (β rhythm), 30-50Hz (γ rhythm). Given the input EEG representation x, the WaveConv at time sample t is

defined by:

$$x_A(t) = \sum_{r=0}^{R} x(s \times t - k) \times u(r)$$
$$x_D(t) = \sum_{r=0}^{R} x(s \times t - k) \times v(r) \quad (1)$$

where u and v represent a pair of wavelet filters, named approximation filter and detail filter, x_A and x_D refer to approximation coefficients and detail coefficients, respectively. R and s are the kernel size and stride in a WaveConv. In order to attain the five subbands aforementioned, the number of WaveConv layer V is decided by the sampling rate of EEG signals: $V = \lfloor \log_2(f_s) \rfloor - 3$. f_s is the original sampling rate of EEG signals, and $\lfloor \cdot \rfloor$ means the rounding-down operation. The stride and kernel size of the WaveConv are set to 2 and 8 respectively [14], both of which consist with the order of the Daubechies wavelet filter. Since we implement the wavelet decomposition by the means of convolution, x_A and x_D are calculated together in the same convolution layer and then separated. Consequently, the number of output channel is twice as large as the number of input channel, which guarantees the same channel number after the following separating operation. Suppose that input channel number is R, and the $2R$ output channels are separated into approximation coefficients and detail coefficients, which are defined by:

$$x_A = \{x_w(c) | c = 1, 3, \ldots, 2R - 1\}$$
$$x_D = \{x_w(c) | c = 2, 4, \ldots, 2R\} \quad (2)$$

where x_w is the output of each WaveConv, c refers to the channel index. For each input channel, a pair of wavelet filters (u, v) is applied and two output channels are generated. In this way, x_A and x_D are arranged alternatively in the output channels, so they need to be picked alternatively by Eq. (2). Furthermore, we apply a special padding method to transform the WaveConv results into periodic and smooth representations, which alleviates the distortion problem especially in head and tail of signals after the wavelet convolution [14]. Given the 1-D input x_A, the periodic padding is defined by:

$$\tilde{x}_A = x_A(K - {}^{h}/_{2} + 1), \ldots, x_A(K-1) \copyright x_A(0), \ldots,$$
$$x_A(K-1) \copyright x_A(0), \ldots, x_A({}^{h}/_{2} - 2) \quad (3)$$

where $\tilde{x}_A$ represent the padding result of x_A, and © refers to the concatenating operation. K is the length of signals, and h refers to the kernel size of wavelet convolutions. As a result, the MS-Conv block generates parallel groups of primary intra-rhythm spectral representations via the designed WaveConv.

4) The Design of SE-Feature-Fusion Block: From the aforementioned blocks, deep-temporal and multi-spectral features are captured independently. However, although extracting abundant features is proved to be helpful in MI-EEG decoding, irrelevant or redundant information are usually inevitable if directly combing all features [26]. Therefore, in order to effectively fuse the features, a SE-Feature-Fusion block is employed to solve the redundancy problem in feature fusion and emphasize the most discriminative features.

Considering the selection of feature subset can significantly benefit the performance of the motor imagery EEG classification [27], we introduce a variant of SE operation, termed squeeze-excitation-convolution (SEC) unit, to enable the model to emphasize the most informative features in the feature fusion process. Fig. 2(b) depicts the structures of the SEC-based SE-Feature-Fusion block. For the input feature maps, the SEC unit recalibrates the features by performing a "squeeze" operation, which aggregates feature maps across time dimension to produce a descriptor [17]. Specifically, in order to use the information of channel dependencies, the "squeeze" operation firstly performs global-average pooling on the input feature maps to squeeze global information, which achieves the purpose of generating channel-wise statistic features. Formally, the statistic output calculated by global-average pooling are defined by:

$$z_{sq} = \frac{1}{T} \sum_{t=1}^{T} x_s(m, c, t) \quad (4)$$

where $x_s \in R^{m \times c \times t}$ is the input feature maps, T is the length of the time samples, and z_{sq} is the squeezing result. After extracting channel-wise information from the "squeeze" operation, the following "excitation" operation is applied to fully utilize the channel dependencies. This is achieved by two fully-connection (FC) layers, the ELU function and the Softmax function. Then we apply the generated channel information to the input features by a multiplication between the learned channel weights and feature maps. The output of the SE operation is denoted by:

$$z_{se} = \sigma\left(W_1 \varepsilon\left(W_2 z_{sq}\right)\right) x_s \quad (5)$$

where W_1 and W_2 refer to the weights of first and second fully-connection respectively, $\varepsilon(\cdot)$ is the ELU function, $\sigma(\cdot)$ is the Softmax function, z_{se} is the multiplying result of the SE operation. Compared with the original excitation operation [17], we choose Softmax instead of Sigmoid, since the size information between the input vectors are preserved after the Softmax. Additionally, the reduction ratio r in SEC unit is set to 8. All these nonlinear and hyperparameter settings will be discussed in Section-III. After the channel-wise multiplying, a 1-D convolution is applied to each channel to further extract significant features and match the size of the temporal and spectral feature maps. An adaptive pooling is then followed to reduce feature shapes and generate coarser representations. Finally, the output of the SEC unit is obtained by:

$$x_{sec} = \frac{1}{\tilde{T}} \sum_{t=1}^{\tilde{T}} Conv(z_{se})(m, c, t) \quad (6)$$

where x_{sec} is the final output of the SEC unit, $Conv(\cdot)$ refers to the convolution, $\tilde{T}$ is the length of time samples after the convolution, and the summation represents the adaptive pooling. Finally, by concatenating the outputs of the two SEC units, the SE-Feature-Fusion block generates fused feature maps, which is shown in Fig. 2(b).

5) The Design of Classification Block: Based on the previously fused temporal-spectral features, the classification block is designed to give the final decoding results. First, all feature maps are flattened into 1-D feature vector. The vector is then fed into the fully-connection layer. Finally, the Softmax function transforms the outputs into classification probabilities.

Consequently, the label with the max probability is considered as the final decoding result.

In summary, the optimizing procedure of the proposed decoding framework is shown in ***Algorithm 1***.

Algorithm 1: The Optimization Steps of the Proposed TS- SEFFNet Method

Input: MI-EEG training set E_{train}, validation set E_{val}, learning rate ξ, early stop patience τ, regularization weight λ, the proposed TS-SEFFNet $Net(\cdot)$;
Output: The decoding results O of MI-EEG;
Initialize data X_i in one batch with $i = 1, 2, \ldots, N$;
Initialize parameters in the proposed TS-SEFFNet as $\Theta^{(0)}$, $\xi = 1 \times 10^{-3}$, $\lambda = 1 \times 10^{-2}$, $q = 0$, $e = 0$, $\tau = 160$, $\kappa_{max} = 0$, $\text{epoch}_{\max} = 1000$;
while $q < \text{epoch}_{\max}$ **and** $e \leq \tau$
  q ++;
  Generate conditional probability $p_j = Net(\Theta^{(q)}, X_i)$;
  Calculate loss $J^{(q)}$ on E_{train} by Eq. (7);
  Calculate the gradient $g = \nabla J^{(q)}$;
  Calculate average accuracy $\kappa^{(q)}$ on E_{val};
  if $\kappa^{(q)} \leq \kappa_{max}$
    e ++;
  else
    $e \leftarrow 0$;
    $\kappa_{max} \leftarrow \kappa^{(q)}$;
  Update parameters $\Theta^{(q+1)} \leftarrow \Theta^{(q)} - \xi \times g$;
end while
Combine E_{train} and E_{val} into one set E;
while $q < 2\text{epoch}_{\max}$ **and** $\kappa^{(q)} \leq \kappa_{max}$
  q ++;
  Generate conditional probability $p_j = Net(X_i)$;
  Calculate loss $J^{(q)}$ on E by Eq. (7);
  Calculate the gradient $g = \nabla J^{(q)}$;
  Calculate average accuracy $\kappa^{(q)}$ on E;
  Update parameters $\Theta^{(q+1)} \leftarrow \Theta^{(q)} - \xi \times g$;
end while
Get the decoding results $\text{O} = Net\left(\Theta^{(q)}, X_i\right)$.

Specifically, the raw signals are firstly filtered into a third-order Butterworth bandpass filter [28] before fed into the TS-SEFFNet, which helps to minimize artifacts and information with low relevance. Additionally, for two datasets, we adopt an early stopping strategy [28] during the training phase. The original training set is split into two folds (training set and validation set). When the validation accuracy does not increase after the early stop patience τ, the training process is stopped. After that, the second training phase starts from the parameters saved in the first stop. The early stopping generally makes it possible to train on different datasets without deciding the number of epochs manually. Moreover, the early stopping avoids the overfitting problem since the training stops at the best epoch. If the training continues after the best epoch, the model generalization ability probably decreases badly. Finally, the model is trained by minimizing the cross-entropy loss J between model predictions and labels:

$$J = \sum_{j=1}^{N} \sum_{i=1}^{M} -\log\left(p_i\right) \omega\left(y_j = l_i\right) + \lambda \left\|\Theta\right\| \quad (7)$$

where p_j is the j-th conditional probability generated by the model, l_j is the j-th class from label set L, $\omega(\cdot)$ is the indicator function. Θ represents learnable parameters of the model, $\|\cdot\|$ is the regularization item for alleviating the overfitting problem, and λ refers to the trade-off regularization weight. We normalize the initial weights $\Theta^{(0)}$ of the proposed TS-SEFFNet with zero mean and variance of 1. The detail model initialization is given in ***Algorithm 1***.

III. Experimental Results

A. EEG Datasets

The effectiveness of the proposed TS-SEFFNet is evaluated on two public MI-EEG datasets which are described in this section.

1) BCI Competition 2008 IV 2a Dataset: (BCI IV 2a) [29] consists of EEG data from 9 subjects. This BCI paradigm is cue-based, including four different MI tasks (left hand, right hand, feet and tongue). Each subject has two sessions and there are 288 trials (72 for each class) per session. The EEG signals were recorded at 250 Hz by 25 electrodes, and the three EOG electrodes are not used for decoding. In this paper, the proposed TS-SEFFNet is trained on the first session (288 trials) and tested on the second session (288 trials). The raw signals are filtered by third-order Butterworth lowpass filter into 0-38Hz [28] before training and testing, which helps to minimize artifacts and information with low relevance.

2) High Gamma Dataset: (HGD) [28] was recorded by 128 electrodes, including EEG signals from 14 healthy subjects. HGD also consists of four different MI tasks (left hand, right hand, feet and rest). Approximately there are 880 trials for training and 160 trials for testing in each subject, and the signal sampling rate is 500 Hz. In this paper, HGD is pre-processed by following the same steps in [28]. First, the 44 electrodes which cover the motor cortex are selected for MI-EEG decoding. Second, signals are filtered by third-order Butterworth lowpass filter into 0-125Hz. Next, the HGD is resampled to 250Hz, i.e., the same as the BCI IV 2a. Note that resampling for HGD is necessary, which makes it able to use a unified hyperparameter settings for the proposed deep-learning framework for both datasets. The 44 selected electrodes can be found in [28].

For both datasets, EEG signals of each trial were extracted by using the same time window $[-0.5\text{s}, 4\text{s}]$ (relative to the cue onset). In addition, the original training set for each subject in both datasets is divided into ten folds (nine for training and one for validation). Detailed data segmentation can be found in Table I.

B. Evaluation Metrics and Models

In the experiments, we use classification accuracy (ACC) [26], Cohen's kappa coefficient (K) [11], F1-score (F1) [30] and area under curve (AUC) [31] to evaluate the proposed

TABLE I
DATA SEGMENTATION FOR EACH SUBJECT IN TWO DATASETS

Dataset	Training Set	Validation Set	Test Set	Total
BCI IV 2a	259	29	288	576
HGD	792	88	160	1040

TS-SEFFNet, where the Cohen's kappa coefficient is denoted by:

$$\mathrm{K} = \frac{ACC - p_e}{1 - p_e}$$
$$p_e = \frac{\sum_{i=1}^{M} n_{:i} n_{i:}}{N^2} \tag{8}$$

where p_e represents the 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 refers to the class number and N is the sum of all entries in the confusion matrix.

We use three baseline models to make overall comparison with the proposed method. All these models are retested on two datasets in this paper. Brief descriptions of the compared baseline models are given as follows:

1) **Shallow ConvNet [28]**: This is a deep learning model with two simple convolution layers and a mean pooling layer, which has shown its ability in MI-EEG decoding.
2) **Deep ConvNet [28]**: This is a deeper model compared to Shallow ConvNet, using three more convolution layers among time dimension, which is probably more suitable when dealing with larger dataset.
3) **CP-MixedNet [19]**: This model uses multi-scale EEG features generated from several convolution layers, and each layer extracts EEG temporal representation from different scales.

Additionally, for further investigating the importance of each block in our TS-SEFFNet, we propose some simplified models to conduct ablation studies, which are introduced as follows:

1) **Temporal-ConvNet**: This convolutional network includes the spatio-temporal block, the DT-Conv block and the classification block, extracting deep-temporal features only.
2) **Spectral-ConvNet**: This network contains spatio-temporal block, MS-Conv block and classification block, merely extracting multi-spectral features.
3) **TS-ConvNet**: This network involves the spatio-temporal block, the DT-Conv block, the MS-Conv block, and the classification block, which uses naive feature fusion that merely concatenates temporal and spectral features.

C. Overall Performance

The proposed TS-SEFFNet is a more competitive method in the presence of the deep-temporal feature extraction, multi-spectral analysis and effective feature fusion. In this section, in order to evaluate the proposed TS-SEFFNet, we compare our TS-SEFFNet with the above baseline models.

Table II and Table III summarize the decoding results on both datasets by using the proposed TS-SEFFNet and baseline methods. From Table II we can learn that the Shallow ConvNet, Deep ConvNet and CP-MixedNet can achieve relatively high average accuracies on 9 subjects. Especially, our TS-SEFFNet reaches the highest average accuracy of 74.71%, which is 1.79%, 2.72% and 7.54% higher than baseline methods respectively, indicating its ability in learning discriminative deep-temporal features while combining robust multi-spectral representations. Moreover, for the average Kappa value, our method outperforms all compared studies and gains the highest value of 0.663. Meanwhile, our method yields average F1-score of 0.757 and average AUC of 0.922, which are the best among these methods. As a result, the proposed TS-SEFFNet achieves promising results and proves its ability in MI-EEG decoding.

We also test the proposed TS-SEFFNet and baseline methods on HGD to further evaluate their decoding performance. From Table III we can see that our TS-SEFFNet can achieve encouraging results of 93.25% in average accuracy and 0.910 in Kappa value, which are at least 3.71% and 0.049 higher than the compared studies. Additionally, as for F1-score and AUC, the proposed TS-SEFFNet reaches to the highest results of 0.901 and 0.988, which outperform all these methods. The above experimental results on two datasets demonstrate the ability of our TS-SEFFNet for MI-EEG decoding.

D. Performance Depending on the DT-Conv Block

To assess the effectiveness of the DT-Conv block in capturing critical deep-temporal features, we conduct ablation studies on the TS-ConvNet and the Spectral-ConvNet which are two simplified models compared to the proposed TS-SEFFNet, and results are given in Table IV. We can observe that, compared with Spectral-ConvNet which excludes DT-Conv block, the TS-ConvNet reaches higher accuracies on both datasets and gains increases of 8.76% and 14.29%, respectively. Also, the Kappa values show increases of 0.116 and 0.09. The increases clearly demonstrate the effectiveness of the DT-Conv block, which explores deeper temporal information embedded in EEG signals and extracts robust deep-temporal features. Fig. 3 illustrates how we optimize the number of the temporal conv units and verifies its effectiveness by comparing results from different unit numbers. From the average accuracy we can see that the higher performance is achieved when applying 3 units. Reducing the number of blocks leads to decreasing of the classification accuracy, while increasing the unit number results in inadequacy of temporal representation, since EEG feature maps are relatively short in time dimension after 3 temporal conv units. The influence of temporal convolution kernel size is also evaluated by the cross-validation experiments, and Fig. 3 demonstrates that the proposed TS-SEFFNet shows its optimal performance at the size of 11 on both datasets. Additionally, the ROC curves in Fig. 4 reach to the highest AUC value of 0.93 and 0.988 respectively by using 3 temporal conv units. As a result, the above results indicate that DT-Conv block captures essential temporal-domain embeddings effectively.

E. Performance Depending on the MS-Conv Block

In this section, the efficiency of the MS-Conv block is analyzed in ablation studies by comparing the TS-ConvNet

TABLE II
THE OVERALL COMPARISON OF CLASSIFICATION PERFORMANCE ON BCI IV 2A

Subject	Shallow ConvNet				Deep ConvNet				CP-MixedNet				**Our TS-SEFFNet**			
	ACC(%)	K	F1	AUC	ACC(%)	K	F1	AUC	ACC(%)	K	F1	AUC	ACC(%)	K	F1	AUC
A01	80.20	0.736	0.800	0.963	80.90	0.745	0.809	0.955	74.65	0.662	0.744	0.907	**82.29**	**0.764**	**0.830**	**0.959**
A02	**54.86**	**0.398**	**0.549**	**0.803**	52.08	0.361	0.523	0.791	53.47	0.380	0.535	0.785	49.79	0.331	0.450	0.748
A03	**92.01**	**0.894**	**0.920**	**0.992**	84.72	0.796	0.846	0.981	73.26	0.644	0.731	0.916	87.57	0.834	0.899	0.988
A04	59.03	0.454	0.590	0.859	71.18	0.616	0.712	0.888	70.14	0.602	0.694	0.909	**71.74**	**0.623**	**0.729**	**0.916**
A05	66.67	0.556	0.664	0.887	70.49	0.606	0.704	**0.922**	67.36	0.565	0.675	0.893	**70.83**	**0.611**	**0.729**	0.921
A06	54.17	0.389	0.541	0.803	55.56	0.407	0.552	0.825	48.96	0.319	0.488	0.793	**63.75**	**0.517**	**0.652**	**0.864**
A07	**87.50**	**0.833**	**0.875**	**0.988**	69.10	0.588	0.685	0.922	74.31	0.657	0.742	0.925	82.92	0.772	0.874	0.986
A08	79.51	0.727	0.796	**0.955**	**81.94**	**0.759**	**0.819**	0.953	72.92	0.639	0.730	0.926	81.53	0.754	0.808	0.952
A09	**82.29**	**0.764**	0.821	0.956	81.94	0.759	0.819	0.951	69.44	0.593	0.694	0.903	81.94	0.759	**0.838**	**0.964**
Aver	72.92	0.639	0.728	0.912	71.99	0.627	0.719	0.910	67.17	0.562	0.670	0.884	**74.71**	**0.663**	**0.757**	**0.922**

where the bold values indicate the best results.

TABLE III
THE OVERALL COMPARISON OF CLASSIFICATION PERFORMANCE ON HGD

Subject	Shallow ConvNet				Deep ConvNet				CP-MixedNet				**Our TS-SEFFNet**			
	ACC(%)	K	F1	AUC	ACC(%)	K	F1	AUC	ACC(%)	K	F1	AUC	ACC(%)	K	F1	AUC
1	**91.25**	**0.883**	**0.913**	**0.992**	81.88	0.758	0.803	0.984	88.75	0.850	0.885	0.990	90.69	0.876	0.890	0.988
2	85.63	0.808	0.852	0.981	91.88	0.892	**0.918**	0.993	90.00	0.867	0.898	0.989	**93.53**	**0.914**	0.897	**0.995**
3	96.88	0.958	0.968	0.999	93.13	0.908	0.931	0.998	95.63	0.942	0.956	0.997	**98.53**	**0.980**	**0.981**	**0.999**
4	93.13	0.908	0.931	0.988	92.50	0.900	0.924	0.995	91.25	0.883	0.912	0.991	**96.88**	**0.958**	**0.962**	**0.998**
5	95.00	0.933	0.950	0.994	90.63	0.875	0.907	0.995	**95.00**	**0.933**	**0.950**	**0.996**	92.90	0.905	0.910	0.994
6	89.38	0.858	0.893	0.984	93.13	0.908	**0.931**	**0.993**	91.25	0.883	0.912	0.987	**93.53**	**0.914**	0.925	0.992
7	89.94	0.866	0.899	0.979	84.28	0.790	0.842	**0.982**	88.05	0.841	0.880	0.956	**92.40**	**0.899**	**0.899**	0.980
8	87.50	0.833	0.874	0.978	90.80	0.877	0.903	0.984	**93.13**	**0.908**	**0.931**	**0.989**	91.78	0.890	0.894	0.968
9	**97.50**	**0.967**	**0.975**	0.997	96.88	0.958	0.969	0.996	95.00	0.933	0.950	0.998	96.88	0.958	0.969	**0.999**
10	80.63	0.742	0.802	0.951	85.00	0.800	0.843	0.973	88.75	0.850	**0.885**	0.970	**89.88**	**0.865**	0.859	**0.974**
11	72.50	0.633	0.729	0.911	88.13	0.842	0.883	0.980	75.63	0.675	0.761	0.926	**92.78**	**0.904**	**0.896**	**0.989**
12	93.13	0.908	0.931	**0.995**	91.25	0.883	0.912	0.994	93.75	0.917	**0.937**	0.990	**95.40**	**0.939**	0.931	0.991
13	85.53	0.807	0.856	0.965	89.94	0.866	**0.900**	0.989	89.31	0.857	0.892	0.987	**93.03**	**0.907**	0.850	**0.989**
14	83.75	0.783	0.836	0.969	83.75	0.783	**0.838**	0.973	78.13	0.708	0.778	0.948	**87.34**	**0.831**	0.756	**0.977**
Aver	88.69	0.849	0.887	0.977	89.51	0.860	0.893	0.988	89.54	0.861	0.895	0.981	**93.25**	**0.910**	**0.901**	**0.988**

where the bold values indicate the best results.

TABLE IV
ABLATION STUDIES ON TWO DATASETS

Dataset	Methods	ACC(%)	K	F1	AUC
BCI IV 2a	Spectral-ConvNet	65.62	0.542	0.550	0.830
	Temporal-ConvNet	73.80	0.651	0.737	0.915
	TS-ConvNet	74.38	0.658	0.746	0.920
	TS-SEFFNet	**74.71**	**0.663**	**0.757**	**0.922**
HGD	Spectral-ConvNet	78.77	0.717	0.823	0.941
	Temporal-ConvNet	92.45	0.899	0.862	0.980
	TS-ConvNet	93.06	0.907	0.883	0.985
	TS-SEFFNet	**93.25**	**0.910**	**0.901**	**0.988**

where the bold values indicate the proposed method.

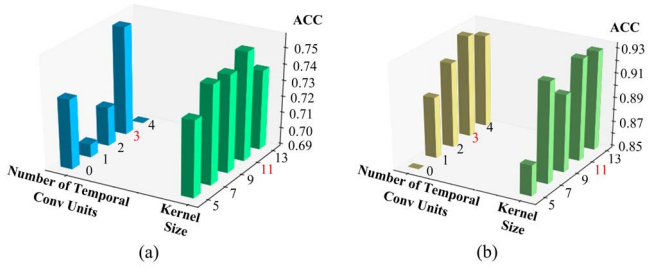

Fig. 3. Accuracy comparison between different kernel sizes and numbers of temporal conv units in DT-Conv block on (a) BCI IV 2a, (b) HGD.

with the Temporal-ConvNet. From Table IV we can learn that the TS-ConvNet outperforms the Temporal-ConvNet in both average accuracy and Kappa value by using the MS-Conv block. This improvement shows that only extracting temporal features from EEG signals may omit important spectral information [32], which proves the efficiency of the MS-Conv block in extracting discriminative multi-spectral features. In addition, we evaluate the impacts of each spectral subband by comparing the classification accuracies on two datasets, and results are shown in Table V. As for applying a single subband, average accuracies range from 72.92% (θ rhythm) to 73.46% (δ rhythm) on BCI IV 2a and from 90.32% (β rhythm) to 92.36% (δ rhythm) on HGD, which indicate that neural activities corresponding to different frequency bands contain different information for MI-EEG decoding. Additionally, when using only δ rhythm representation, average accuracy reaches 73.46% and 92.36% respectively, which are the highest among single rhythm analysis and show the importance of δ rhythm. Moreover, compared with the single subband, the multi-spectral network achieves the highest average accuracy of 74.71% and 93.25% from two datasets, which intuitively shows the effectiveness of multi-spectral analysis.

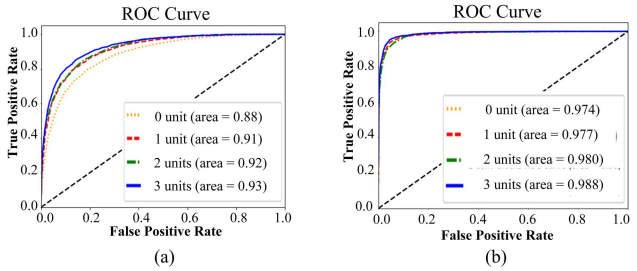

Fig. 4. ROC curve comparison between different numbers of temporal conv units in DT-Conv block on (a) BCI IV 2a, (b) HGD.

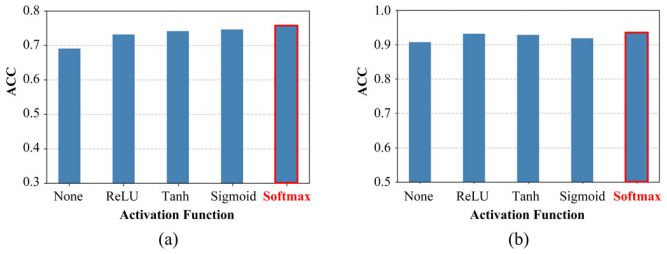

Fig. 5. Accuracy comparison between different activation functions in SEC unit on (a) BCI IV 2a, (b) HGD.

TABLE V
ACCURACY(%) COMPARISON BETWEEN RHYTHMS ON TWO DATASETS

Dataset	Subject	δ	θ	α	β	γ	Overall
BCI IV 2a	A01	78.06	78.13	79.38	80.28	79.86	**82.29**
	A02	**50.49**	49.86	49.24	48.47	50.07	49.79
	A03	85.69	85.56	87.15	84.65	86.63	**87.57**
	A04	**71.94**	71.46	70.97	70.07	71.32	71.74
	A05	**73.68**	71.11	72.64	74.17	71.18	70.83
	A06	60.63	60.21	61.32	59.17	60.35	**63.75**
	A07	76.25	76.04	76.46	77.92	77.43	**82.92**
	A08	81.46	80.21	79.58	79.86	80.49	**81.53**
	A09	82.92	**83.75**	82.71	82.36	82.29	81.94
	Aver	73.46	72.92	73.27	72.99	73.18	**74.71**
HGD	1	86.50	85.25	83.25	82.63	85.75	**90.69**
	2	**94.63**	92.00	93.25	94.50	91.38	93.53
	3	96.63	98.13	96.88	**98.88**	97.63	98.53
	4	96.00	96.25	96.38	**98.13**	97.00	96.88
	5	93.50	94.50	91.50	94.63	**95.75**	92.90
	6	94.63	**95.00**	93.88	95.00	93.88	93.53
	7	90.53	90.05	**94.45**	84.94	89.42	92.40
	8	90.50	90.75	90.13	**93.25**	83.88	91.78
	9	**98.13**	96.88	97.63	96.88	97.50	96.88
	10	88.50	89.63	86.38	84.63	**90.88**	89.88
	11	89.13	90.13	85.75	84.50	91.38	**92.78**
	12	95.38	93.13	93.13	**96.38**	95.00	95.40
	13	**94.08**	88.68	89.53	88.79	90.57	93.03
	14	84.88	74.88	73.88	71.38	79.00	**87.34**
	Aver	92.36	91.09	90.43	90.32	91.36	**93.25**

where the bold values indicate the best results.

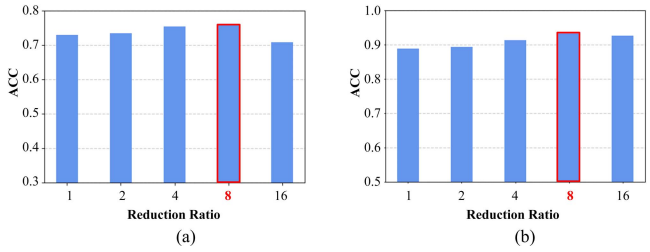

Fig. 6. Accuracy comparison between different reduction ratios in SEC unit on (a) BCI IV 2a, (b) HGD.

TABLE VI
COMPARISON OF PARAMETERS AND COMPUTATIONAL COST

Dataset	Method	Parameter (million)	Inference time (ms)	Decoding time (ms)
BCI IV 2a	Shallow ConvNet[28]	0.047	1.116	36.20
	Deep ConvNet[28]	0.284	1.215	36.28
	CP-MixedNet[19]	0.836	1.450	36.60
	Our method	**0.282**	**1.168**	**36.22**
HGD	Shallow ConvNet[28]	0.083	1.401	20.99
	Deep ConvNet[28]	0.298	1.961	31.70
	CP-MixedNet[19]	0.842	2.420	43.73
	Our method	**0.296**	**1.915**	**27.24**

where the bold values indicate the proposed method.

F. Performance Based on the SE-Feature-Fusion Block

Apart from the spectral information analysis, the SE-Feature-Fusion block also shows great effects on the decoding results. We first consider the choice of nonlinearity in the proposed SEC unit. For the activation function employed, four options including ReLU, Tanh, Sigmoid and Softmax are compared in Fig. 5 on two datasets. We can see that using ReLU gets the worst average accuracy, while replacing it with Tanh or Sigmoid gains higher decoding results. The highest accuracy is obtained with Softmax function, and this suggests that the construction of the excitation operator is significant [17]. Moreover, for the reduction ratio r in the SEC unit, we set values ranging from 1 to 16 to find the best hyperparameter. Fig. 6 shows the average accuracies on two datasets with different reduction ratios, and we can observe that our proposed TS-SEFFNet reaches the best classification accuracy with the reduction ratio 8.

In order to evaluate the ability of the SE-Feature-Fusion block, the proposed TS-SEFFNet is compared with TS-ConvNet which excludes the SE-Feature Fusion block. In Table IV, the proposed TS-SEFFNet reaches average accuracies of 74.71% and 93.25% on two datasets, which are 0.33% and 0.19% higher than the TS-ConvNet. The improvements manifest that compared with naive concatenation method, the SE-Feature-Fusion block indeed helps to alleviate the heterogeneity of redundant features and boost the most discriminative features. Moreover, the experiment results indicate that merely extracting features from DT-Conv block and MS-Conv block is not enough for the MI-EEG decoding, and the effective fusion significantly increases the decoding performance. Therefore, we can learn that the SE-Feature-Fusion block contributes to our proposed model in contrast to the naive feature fusion.

IV. DISCUSSIONS

A. Efficacy of Model Compactness

Commonly, a model is expected to have as fewer learnable parameters as possible to ensure its robust generalizability [28]. In order to further evaluate the computational

TABLE VII
COMPARISON WITH THE STATE-OF-THE-ART METHODS

Dataset	Methods	Year	ACC(%)	K
BCI IV 2a	Deep ConvNet[28]	2017	70.90	0.612
	C2CM[15]	2018	74.46	0.659
	CP-MixedNet[19]	2019	73.20	0.643
	RA-MDRM[33]	2020	70.91	0.612
	EA-CSP-LDA[33]	2020	73.53	0.647
	FBSF-TSCNN[11]	2020	72.00	0.627
	Our method	**2021**	**74.71**	**0.663**
HGD	Deep ConvNet[28]	2017	92.50	0.900
	Hybrid Net[28]	2017	91.80	0.891
	Residual Net[28]	2019	88.90	0.852
	FBCSP[19]	2019	90.90	0.879
	CP-MixedNet[19]	2019	93.00	0.907
	Our method	**2021**	**93.25**	**0.910**

where the bold values indicate the proposed method.

TABLE VIII
10-FOLD CROSS VALIDATION COMPARISON ON BCI IV 2A

Dataset	Methods	Year	ACC(%)	K
BCI IV 2a	PSO-Rough Set[5]	2016	80.99	0.743
	Brain Network[8]	2017	79.70	0.730
	3D-CNN[12]	2019	75.02	0.644
	Discriminative Feature[9]	2021	81.85	0.758
	Our method	**2021**	**84.49**	**0.794**

where the bold values indicate the proposed method.

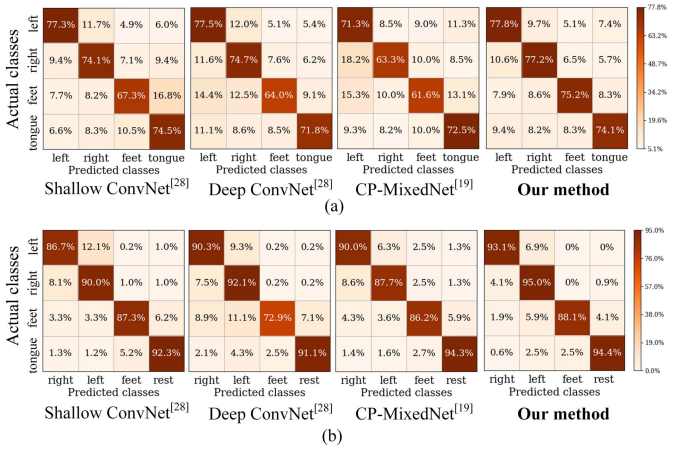

Fig. 7. The confusion matrixes on both (a) BCI IV 2a, (b) HGD.

complexity of the proposed network, we compute the number of parameters in our TS-SEFFNet and compare it with baseline methods, and the results are listed in Table VI. The proposed TS-SEFFNet contains approximately 2.82×10^5 parameters, which is similar to the Deep ConvNet with 2.84×10^5 parameters. Moreover, compared to the CP-MixedNet with parameters up to 8.36×10^5 [19], our TS-SEFFNet is compact enough for MI-EEG decoding task, which can reduces the model complexity by approximately 66.63%.

Next, we test the model inference time, which is defined as the duration when the trained model gives a classification for one MI-EEG trial, and the results are shown in Table VI. The computations are implemented on a standard computer with a 2.6 GHz processor and 8 GB RAM. The training and testing of the deep learning model are performed on Nvidia Tesla V100 GPU with 32 GB memory. The inference time is the average result calculated on all subjects of each dataset. We can see that the CP-MixedNet and the Deep ConvNet consume more time when making inferences, while our TS-SEFFNet is relatively fast among these methods.

From the number of parameters and the inference time, we can conclude that the proposed TS-SEFFNet achieves an appropriate tradeoff between complexity and performance. The compactness of our TS-SEFFNet is highly relied on the use of WaveConv layer, which includes no learnable parameters and has smaller filter size in convolution operations. Also, the max pooling layers in DT-Conv block and global pooling layers in SEC unit decrease the feature map length while retaining the discriminative features. Additionally, we record the decoding time of each method, which starts from raw EEG and ends with decoding results. The decoding time of our TS-SEFFNet on one EEG trial are 36.22ms and 27.24ms for two datasets respectively, which are within 1s and suitable enough for a real-time BCI system [1].

B. Comparison of Different Decoding Methods Reported for BCI IV 2a and HGD

Table VII makes comparison between decoding results reported by recent studies on both datasets. All these methods train their models on the first session (288 trials for BCI IV 2a and 880 trials for HGD) and test on the second session (288 trials for BCI IV 2a and 160 trials for HGD) for each subject. The detailed data segmentation for two datasets can be found in Table I.

From Table VII we can learn that the proposed TS-SEFFNet illustrates better classification performance than most of these methods for BCI IV 2a dataset. Compared with methods in [15] and [28] which designed CNNs to capture temporal or spatial features, our method uses both deep-temporal and multi-spectral features for MI-EEG decoding and gains accuracy increases of 0.25% and 3.81% respectively. Chen *et al.* [11] built a temporal-spatial CNN which explored spatio-temporal features of EEG but resulted in suboptimal performance compared with our method. This is probably due to the lack of multi-spectral feature extraction in their approaches. Moreover, in comparison with EA-CSP-LDA [33] which tried to alleviate spatial discrepancies of EEG trials, our method emphasizes the importance of deep temporal features and reaches 1.18% higher in accuracy. Table VII also lists comparison between our TS-SEFFNet and other methods on HGD. For example, the Hybrid Net [28] combines shallow and deep convolution layers for classification and the CP-MixedNet [19] applies CNN to extract multi-scale temporal features. In contrast, our method considers both deep-temporal and multi-spectral features simultaneously, which gains accuracy increases of 1.6% and 0.4%. Consequently, the higher classification results demonstrate that the proposed TS-SEFFNet can decode multi-task MI-EEG more accurately and effectively.

Additionally, for BCI IV 2a dataset, some recent studies [5], [8], [9], [12] report results by using 10-fold cross-validation on merged data (576 trials) for each subject. Therefore, in order to make a fair comparison with these conventional methods, Table VIII gives the comparison between

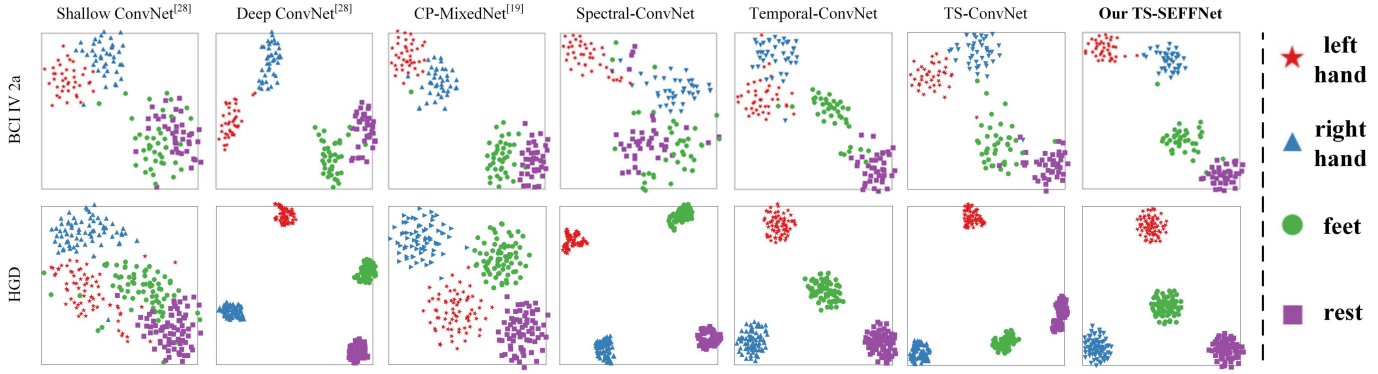

Fig. 8. The t-SNE visualization in 2-D embedding space of features learned by different methods from the third subject in BCI IV 2a and HGD.

the proposed TS-SEFFNet and recent studies on BCI IV 2a by training on merged data (576 trials) for each subject. Compared with Zhao *et al.* [12] who designed multi-branch 3D CNNs to generate spatial representation from EEG, our TS-SEFFNet captures not only spatio-temporal features but also multi-spectral features for MI-EEG decoding and gains accuracy increase of 9.47%. Furthermore, Kumar *et al.* [5] utilized a feature selection algorithm and Ai *et al.* [8] used a naive concatenation for feature fusion. Compared with these two methods, the proposed TS-SEFFNet uses an attention-based feature fusion method to emphasize the most discriminative feature maps and fuse the captured features effectively, which achieves accuracy increases of 3.50% and 4.79%, respectively.

Note that for HGD, most studies only reported their results by training on the first session (880 trials) and testing on the second session (160 trials), and little literature merges HGD from two sessions into one session (1040 trials) for 10-fold cross validation [19], [28], so the experiment results of merged data on HGD are not provided in this paper.

C. Feature Discrimination Discussion

We exploit the effectiveness of the feature extracted by the proposed TS-SEFFNet using confusion matrix, and experiment results are presented in Fig. 7. We can see that our TS-SEFFNet gains obvious accuracy improvements in four MI tasks on both datasets, with a maximum increase of 13.9% in "left hand" task (BCI IV 2a) and increase of 15.2% in "feet" task (HGD).

In order to further investigate the discrimination of the features extracted by our TS-SEFFNet, the t-SNE is utilized to get visualization of the learned features [34]. The t-SNE visualizes the extracted EEG features into a 2-D embedding dimension, which is shown in Fig. 8. Compared with Shallow ConvNet, Deep ConvNet and CP-MixedNet which omit the spectral features, our TS-SEFFNet implements multi-spectral feature extraction and captures more separable features from MI-EEG. In addition, the feature visualizations from the Spectral-ConvNet, the Temporal-ConvNet and the TS-ConvNet are relatively ambiguous in contrast to our TS-SEFFNet, since the proposed model is able to extract both temporal and spectral features from EEG signals. Moreover, with the SE-Feature-Fusion block, the proposed TS-SEFFNet generates more separable features than the TS-ConvNet, which can efficiently distinguish different types of MI-EEG signals. Therefore, we can clearly see that our TS-SEFFNet extracts the most discriminative EEG features, indicating the best decoding performance.

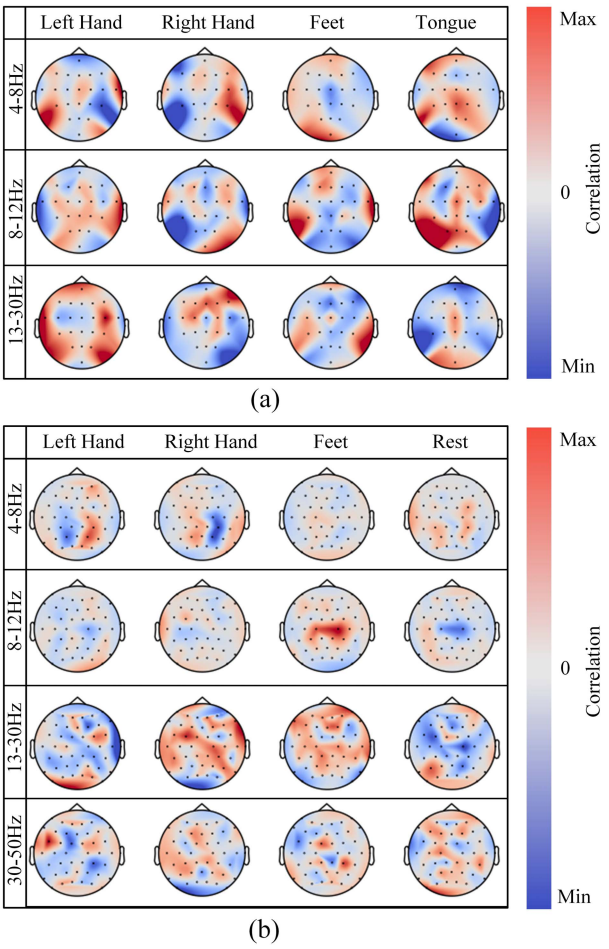

Fig. 9. The network-prediction correlation maps for our TS-SEFFNet on (a) BCI IV 2a (subject A03), (b) HGD (subject 3).

D. Channel Correlation Analysis

We compute the network-prediction correlation on the proposed TS-SEFFNet in different frequency bands, which is visualized in Fig. 9. The visualization on the scalp maps presents spatial distributions expected for MI tasks in the

corresponding rhythms, directly reflecting how the proposed network behaves when making inferences on MI-EEG data. For example, in Fig. 9(a), the positive correlation region on the scalp of "left hand" task (4-8Hz) indicates that model predictions are positively correlated to these corresponding electrodes and frequency bands. Our TS-SEFFNet increases its classification probability on "left hand" when the positive-correlation electrodes gain higher amplitude values, where these results are in line with [28]. In summary, the channel correlation analysis shows the relationships between EEG channels with different motor tasks in the corresponding frequency bands. It is useful to match the feature distributions learned by our TS-SEFFNet with different MI tasks in different frequency bands, which reveals potential relationships between body movements and their associated changes in the brain activities.

E. Limitations and Future Directions

Although the proposed TS-SEFFNet achieves robust decoding results, our present work still suffers from several limitations. First, EEG electrodes are selected manually (i.e. HGD), which probably omits the spatial information of EEG sensors [35], and this neglecting of spatial dependencies in EEG signals may leads to the suboptimal decoding performance [36]. Therefore, our important future work is to apply adaptively-selecting method which focuses on the most discriminative channels. Second, although our method shows the effectiveness in subject-specific MI-EEG decoding scenario, it cannot be used for cross-subject MI-EEG decoding task directly, where the model is applied to a completely new subject after training. This is mainly because that our method is not able to handle the drifting in distributions between data from different subjects [37]. Therefore, the transfer learning [33], [37], [38] will be considered in our feature work to improve our TS-SEFFNet method.

V. Conclusion

In this paper, we propose a novel temporal-spectral-based squeeze-and-excitation feature fusion network (TS-SEFFNet) for MI-EEG decoding. Specifically, our TS-SEFFNet first extracts preliminary spatio-temporal embeddings from raw EEG signals via the spatio-temporal block. Next, the DT-Conv block learns discriminative high-level EEG information through a series of temporal conv units. Meanwhile, the proposed MS-Conv block is adopted to capture essential spectral representations, which integrates multi-level spectral analysis into the end-to-end model. Moreover, in order to effectively fuse the extracted features, the SE-Feature-Fusion block is further employed to emphasize the most discriminative features and reduce redundant feature information. We conduct experiments on two public MI-EEG datasets to evaluate the effectiveness and generalization of the proposed TS-SEFFNet. Our method shows promising results in accuracy, Kappa value, F1-score and AUC value compared with other methods. The experimental results confirm that the proposed method is able to decode MI-EEG efficiently, which can be regarded as a powerful tool for MI-EEG based BCIs.

References

[1] 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.

[2] Y. Zhang, C. S. Nam, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Temporally constrained sparse group spatial patterns for motor imagery BCI," *IEEE Trans. Cybern.*, vol. 49, no. 9, pp. 3322–3332, Sep. 2019.

[3] 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.

[4] R. Zhang, Q. Zong, L. Dou, X. Zhao, Y. Tang, and Z. Li, "Hybrid deep neural network using transfer learning for EEG motor imagery decoding," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102144.

[5] S. U. Kumar and H. H. Inbarani, "PSO-based feature selection and neighborhood rough set-based classification for BCI multiclass motor imagery task," *Neural Comput. Appl.*, vol. 28, no. 11, pp. 3239–3258, 2016.

[6] B. Blankertz, G. Dornhege, M. Krauledat, K. R. Muller, and G. Curio, "The non-invasive Berlin brain–computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539–550, Aug. 2007.

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

[8] Q. Ai *et al.*, "Feature extraction of four-class motor imagery EEG signals based on functional brain network," *J. Neural Eng.*, vol. 16, no. 2, Apr. 2019, Art. no. 026032.

[9] 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. [Online]. Available: https://ieeexplore.ieee.org/document/9327487/

[10] S. Saha, K. I. U. Ahmed, R. Mostafa, L. Hadjileontiadis, and A. Khandoker, "Evidence of variabilities in EEG dynamics during motor imagery-based multiclass brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 371–382, Feb. 2018.

[11] J. Chen, Z. Yu, Z. Gu, and Y. Li, "Deep temporal-spatial feature learning for motor imagery-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2356–2366, Nov. 2020.

[12] 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.

[13] R. Zhang, Q. Zong, L. Dou, and X. Zhao, "A novel hybrid deep learning scheme for four-class motor imagery classification," *J. Neural Eng.*, vol. 16, no. 6, Oct. 2019, Art. no. 066004.

[14] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, "Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze- and-excitation network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 782–794, Apr. 2020.

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

[16] H. Wang *et al.*, "An approach of one-vs-rest filter bank common spatial pattern and spiking neural networks for multiple motor imagery decoding," *IEEE Access*, vol. 8, pp. 86850–86861, 2020.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze- and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[18] D. Freer and G.-Z. Yang, "Data augmentation for self-paced motor imagery classification with C-LSTM," *J. Neural Eng.*, vol. 17, no. 1, Jan. 2020, Art. no. 016041.

[19] 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.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutkever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jan. 2014.

[22] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 4, pp. 317–326, Aug. 2008.

[23] Y. Zhang, B. Liu, X. Ji, and D. Huang, "Classification of EEG signals based on autoregressive model and wavelet packet decomposition," *Neural Process. Lett.*, vol. 45, no. 2, pp. 365–378, Apr. 2016.

[24] Y. Li, X. D. Wang, M. L. Luo, K. Li, X. F. Yang, and Q. Guo, "Epileptic seizure classification of EEGs using time-frequency analysis based multiscale radial basis functions," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 386–397, Mar. 2018.

[25] Y. Li, W.-G. Cui, H. Huang, Y.-Z. Guo, K. Li, and T. Tan, "Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the Fisher vector approach," *Knowl.-Based Syst.*, vol. 164, pp. 96–106, Jan. 2019.

[26] Y. Li, J. Liu, Z. Tang, and B. Lei, "Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for MCI identification," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2818–2830, Sep. 2020.

[27] J. Wang, Z. Feng, X. Ren, N. Lu, J. Luo, and L. Sun, "Feature subset and time segment selection for the classification of EEG data based motor imagery," *Biomed. Signal Process. Control*, vol. 61, Aug. 2020, Art. no. 102026.

[28] 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.

[29] M. Tangermann *et al.*, "Review of the BCI competition IV," *Frontiers Neurosci.*, vol. 6, no. 1, p. 55, 2012.

[30] S. Ren, W. Wang, Z.-G. Hou, X. Liang, J. Wang, and W. Shi, "Enhanced motor imagery based brain-computer interface via FES and VR for lower limbs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 8, pp. 1846–1855, Aug. 2020.

[31] H. Li *et al.*, "Massage therapy's effectiveness on the decoding EEG rhythms of left/right motor imagery and motion execution in patients with skeletal muscle pain," *IEEE J. Transl. Eng. Health Med.*, vol. 9, Feb. 2021, Art. no. 2100320.

[32] L. Wang *et al.*, "Automatic epileptic seizure detection in EEG signals using multi-domain feature extraction and nonlinear analysis," *Entropy*, vol. 19, no. 6, p. 222, May 2017.

[33] H. He and D. Wu, "Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 399–410, Feb. 2020.

[34] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, pp. 3221–3245, Jan. 2014.

[35] D. Zhang, L. Yao, K. Chen, S. Wang, P. D. Haghighi, and C. Sullivan, "A graph-based hierarchical attention model for movement intention detection from EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2247–2253, Nov. 2019.

[36] Y. Li, Y. Liu, Y.-Z. Guo, X.-F. Liao, B. Hu, and T. Yu, "Spatio-temporal-spectral hierarchical graph convolutional network with semisupervised active learning for patient-specific seizure prediction," *IEEE Trans. Cybern.*, early access, May 25, 2021, doi: 10.1109/TCYB.2021.3071860.

[37] X. Zhao, J. Zhao, W. Cai, and S. Wu, "Transferring common spatial filters with semi-supervised learning for zero-training motor imagery brain-computer interface," *IEEE Access*, vol. 7, pp. 58120–58130, 2019.

[38] A. M. Azab, L. Mihaylova, K. K. Ang, and M. Arvaneh, "Weighted transfer learning for improving motor imagery-based brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 7, pp. 1352–1359, Jul. 2019.