Multi-Attention Adaptation Network for Motor Imagery Recognition

Peiyin Chen, Zhongke Gao, *Senior Member, IEEE,* Miaomiao Yin, Jialing Wu, Kai Ma, Celso Grebogi

***Abstract*—Brain computer interface (BCI) based on motor imagery Electroencephalogram (EEG) has been widely used in various applications. Despite the previous efforts, the remained major challenges are effective feature extraction and time-consuming calibration procedure. To address these issues, a novel Multi-Attention Adaptation Network integrating multiple attentions mechanism and transfer learning is proposed to classify the EEG signals. Firstly, the multi-attention layer is introduced to automatically capture the dominant brain regions relevant to mental tasks without incorporating any prior knowledge about the physiology. Then, a multi-attention convolutional neural network is employed to extract deep representation from raw EEG signals. Especially, a domain discriminator is applied to deep representation to reduce the differences between sessions for target subjects. The extensive experiments are conducted on three public EEG datasets (Dataset IIa and IIb of BCI Competition IV, High Gamma dataset), achieving the competitive performance with average classification accuracy of 81.48%, 82.54% and 93.97%, respectively. All the results outperform the state-of-the-art algorithms demonstrate the effectiveness and robustness of the proposed method. Importantly, we confirm that it is easier and more appropriate to transfer the information from local brain regions than from the whole brain. This enhances the transfer ability of deep features and hence it improves the performance of BCI systems.**

***Index Terms*—Motor imagery (MI), transfer learning, multiple attentions mechanism, electroencephalogram (EEG), brain-computer interface (BCI).**

I. Introduction

As a novel interaction, the brain-computer interfaces (BCIs) system allows users to send specific commands to control external auxiliary devices by translating their neuronal activities [1], [2]. Since the non-invasiveness and high temporal resolution of Electroencephalogram (EEG), EEG-based BCIs have drawn great attention and have been widely applied to various fields, including driver fatigue detection [3], [4], emotion recognition [5], [6], entertainment for healthy users [7], [8], and others [9], [10].

According to different brain activity mechanisms, EEG-based evoked patterns are generally divided into four categories: sensorimotor rhythms (SMR), steady-state visual evoked potential (SSVEP), event-related potential (ERP) and slow cortical potential (SCP). Motor imagery (MI) associated with SMR is one of the prevalent BCI paradigms. During executing imaginary motor movements, the amplitude of μ (7-13Hz) and β (13-30Hz) rhythms would be supressed or enhanced. These phenomena are called event-related desynchronization (ERD) and event-related synchronization (ERS), respectively [11]. Moreover, the MI-BCIs system does not require external stimulation and interacts with the outside world through the spontaneously controlled EEG. Motor imagery can be used as a means to activate the motor neural network. It can be applied to any stage of stroke to improve the motor function of stroke patients and does not depend on the patient's residual function. It is closely related to the patient's active movement, so the MI-BCIs system has been successfully applied to stroke rehabilitation [12].

In recent years, a large number of methods have been applied to motor imagery EEG signals decoding. These methods are mainly divided into two categories: machine learning methods and deep learning methods. Machine learning methods generally include two stages: feature extraction and classification. In BCIs, the commonly used features include time domain features, frequency domain features, time-frequency joint features and spatial domain features. There are four mainly used methods for extracting time-frequency joint features: short-time Fourier transform (STFT), discrete wavelet transform (DWT), Gabor transform and Wigner-Ville distribution [13]. In addition to being able to select different time resolutions and frequency resolutions, these methods are also able to automatically adjust the time domain resolution and frequency domain resolution to suit the characteristics of the signals. In terms of spatial domain, the common spatial pattern (CSP) is the most typical feature extraction method and is widely used [14]–[16]. The core of the CSP algorithm is to find a set of weight vectors, and use those vectors to spatially project the observed signal so that the difference between the signal variances before and after projection is the largest. Subsequently, more CSP enhancement algorithms were produced, such as filter bank common spatial pattern (FBCSP) [17] and filter bank regularized common spatial pattern (FBRCSP) [18]. Recently, frameworks based

Manuscript received ** **, ****; accepted ** **, ****. This work was supported in part by the National Natural Science Foundation of China under Grants Nos. 61873181 and 61922062 (*Corresponding author: Zhongke Gao, Jialing Wu*).

P. Chen is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: pychen@tju.edu.cn).

Z. Gao is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: zhongkegao@tju.edu.cn).

M. Yin is with Department of Rehabilitation, Tianjin Huanhu Hospital, Tianjin 300350, China (e-mail: yinmiao198801@163.com).

J. Wu is with Department of Neurorehabilitation and Neurology, Tianjin Huanhu Hospital, Tianjin Key Laboratory of Cerebral Vascular and Neurodegenerative Diseases, Tianjin Neurosurgical Institute, Tianjin 300350, China (e-mail: wywjl2009@hotmail.com).

K. Ma is with the Tencent Jarvis Lab, Malata Building, 9998 Shennan Avenue, Shenzhen, Guangdong Province, 518057, China (e-mail:kylekma@tencent.com).

C. Grebogi is with Institute for Complex Systems and Mathematical Biology, King's College, University of Aberdeen, Aberdeen AB24 3UE, UK (e-mail: grebogi@abdn.ac.uk).

on Riemannian Geometry are proposed to extract manifold topological features from symmetric positive definite matrices with the purpose to achieve classification [19], [20]. When classifying the extracted features, the following classifiers are generally used: linear discriminant analysis (LDA), support vector machines (SVM), non-linear Bayesian classifier, nearest neighbor classifier and multiple classifiers, the latter two are used in combination [21]. The classification accuracy of these methods depends to a large extent on handcrafted features. Due to the low signal-to-noise ratio and non-stationary randomness of the EEG signals, the large difference in cross sessions and subjects, it is hard to manually extract robust EEG signal features, resulting in poor classification performance.

Deep learning is a prominent technique for fully mining data and extracting more advanced feature representations, which has been applied to various fields [22], [23]. Specially, various deep learning models have been successfully used to decode EEG signals and achieve good results: Lu *et al.* [24] proposed a frequential deep belief network (FDBN) which is formed by stacking three restricted Boltzmann machines (RBMs) and an extra output layer. These RBMs are trained with the frequency domain features of EEG signals, which are obtained by fast Fourier transform and wavelet packet decomposition. Gao *et al.* [6] designed simple convolutional neural network (ConvNet) based on coincidence filtering, using fewer parameters to tune yielded higher training efficiency for EEG-based classification. Zhao *et al.* [25] first converted the EEG signals to a sequence of 2D array, then transformed it into a 3D representation; they proposed a multi-branch 3D convolutional neural network (3D ConvNet) which can preserve not only temporal features but also spatial ones. Gao *et al.* [26] develop a framework combining recurrence plots and convolutional neural network for fatigue driving recognition.These methods have achieved more effective EEG feature extraction and pattern classification with higher accuracy. However, there are some problems when confronting with brain–computer interfaces: (1) there are fewer training samples in practice and (2) how to leverage data from previous sessions to prevent a calibration for new sessions of the same user. The standard approach in BCIs is re-training a classifier with a series of calibration trials at the beginning of each experimental session. However, such time-consuming procedure is clearly suboptimal, because it does not utilize any information from past experiments.

More recently, transfer learning has been widely applied in many EEG-based BCI studies to address the above challenges. Li $et\ al.$ [27] proposed the bi-hemisphere domain adaptation network (BiDANN) for EEG emotion recognition. In BiDANN, the feature extractor and domain discriminator are introduced to learn a share feature representation space and to reduce the distribution discrepancy between different subjects. Then the classifier trained on EEG signals from source subjects can predict the EEG signals from target subject more accurately. Jeon $et\ al.$ [28] introduced a source selection to choose similar subjects as souce domain by estimating their power spectral density in resting-state EEG signals; then the EEG signals from both source and target domain are utilized to jointly train a deep network. These studies enrich the applications of transfer learning in BCI systems. However, we still encounter challenges in domain adaptation for EEG decoding, that is, (1) the learning problem of share representations space cross source and target domains and (2) the reduction of domain discrepancy.

Additionally, many studies are devoted to optimize the number of EEG channels by searching the maximums of spatial pattern vectors in scalp mappings. For instance, ICA-based methods [29] convert EEG signals into time-frequency maps and scalp maps, then the relevant components task are chosen by visual selection. CSP-based methods [30] consider the spatial pattern as EEG source distribution vectors, and select the channels corresponding to the maximum coefficients of distribution vectors as the optimal channels. SVM-based methods [31] aim at finding a good estimation for the regularization parameter of the objective function, which employed the Fisher Criterion, Zero-Norm Optimization and Recursive Feature Elimination as the estimation criteria. Then the channel scores are ranked during training procedure, and the best ranked channels can be considered as the dominant channels. However, these methods either require an explicit prior knowledge about the physiological processes underlying mental tasks or it is only a preprocessing technique that is difficult to be incorporated in an end-to-end optimization paradigm. Therefore, it presents a major challenge to locate automatically critical channels relevant to mental tasks during training.

In this paper, we propose a multi-attention adaptation network (MAAN) to tackle the aforementioned challenges in BCI applications. We assume that source domain and target domain share the common feature spaces, and that the automatic location of dominant channels during training does boost the capability to extract more discriminative and transferable features. Based on such assumption, the transfer learning technique and multiple attentions mechanism are employed in our framework. Firstly, we design a multi-attention layer with multiple learnable kernels to automatically target the dominant brain regions during training. Then a multi-attention ConvNet incorporating the multi-attention layer is built as the feature extractor to learn discriminative features from EEG signals. Next, these features are fed into both a classifier and a domain discriminator. With the adversarial learning between classifier and domain discriminator, the feature extractor tries capturing the domain-invariant features by minimizing the overall loss. We evaluate the performance of our proposed EEG decoding architecture on three public motor imagery EEG datasets, all the results demonstrate that the proposed MAAN framework achieves superior classification accuracy than other state-of-the-art methods. Moreover, it is shown that our method is capable of locating the critical channel without any prior knowledge for channel selection.

The major contributions of this paper are summarized as follows: We introduce the multiple attentions mechanism to dynamically estimate the channel-wise importance during training, which can be easily implemented for end-to-end learning and is compatible with all EEG-based tasks. The novel multi-attention adaptation network is proposed to learn domain-invariant and discriminative features, which can

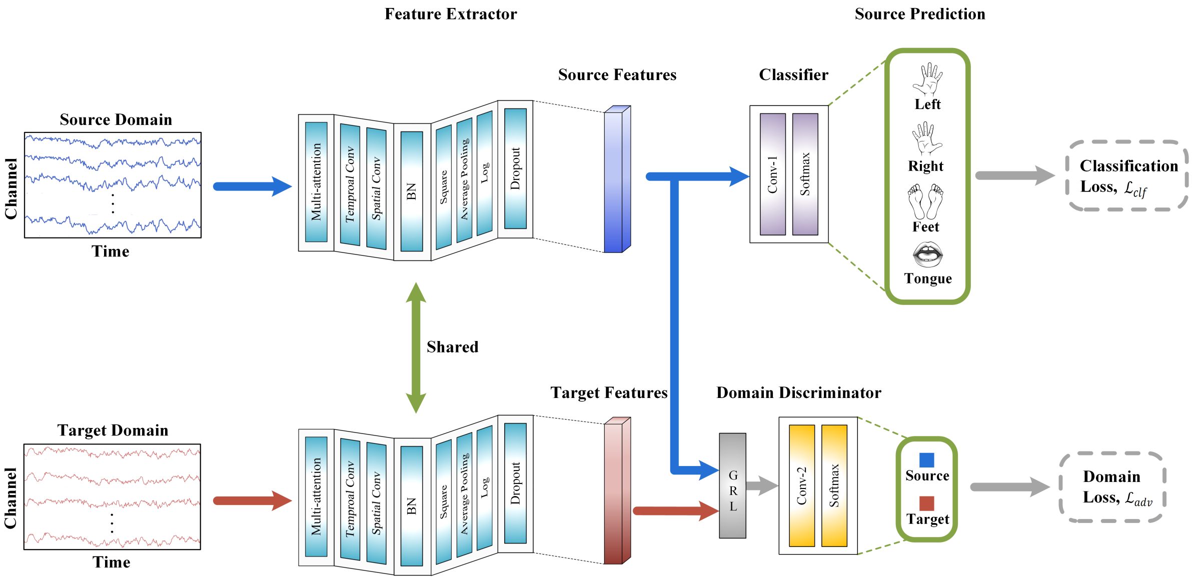


Fig. 1. The architecture of the proposed MAAN model, which consists of three components: feature extractor, classifier and domain discriminator.

reduce the calibration time within sessions effectively. All the experimental results on three public motor imagery EEG datasets not only show that our models outperform many state-of-the-art deep learning methods, but also confirm that transferring the information from the local dominant brain regions is easier and more realistic.

The remainder of this paper is organized as follows: Section II briefly introduces some notations and preprocessing technique that we consider in this work. Section III presents the components of the proposed method in detail. Section IV describes the experiment and discusses the results on three public motor imagery EEG datasets (Dataset IIa and IIb of BCI competition IV, High Gamma dataset). Lastly, Section V concludes the paper.

II. Preliminary

In this section, we first present the definitions and notations used in this work and introduce the adopted preprocessing technique for EEG signals.

A. Definitions and Notations

The EEG signals collected from a subject are defined as $\{(x_i, y_i)\}_{i=1}^{n}$, where n denotes the number of EEG trials. $x_i \in \mathbb{R}^{C\times T}$ represents an EEG trial with C channels and T sampling points, y_i is the corresponding label.

In this study, we only consider the case of inter-session transfer for the same subject. Specially, the EEG trials from previous session are defined as source domain, and trials from new session are defined as target domain. Usually, we denote the source domain with n_s labeled samples as $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, where x_i^s represents the data drawn from the marginal distribution $P(X_s)$, y_i^s is the label of x_i^s. Similarly, we define $D_t = \{(x_i^t)\}_{i=1}^{n_t}$ as the target domain with n_t unlabeled samples under the marginal distribution $P(X_t)$. In general, it is assumed that the data come from different domains following the similar but different marginal distribution, called *domain shift*. For a classification task, our goal is to train a classifier that leverages the available information from source domain to reduce the domain shift, and make it perform well on the classification of data points for the target domain.

B. Data Preprocessing

Before presenting the details of the network architecture, it is necessary to introduce the preprocessing procedures used in this work. Compared to traditional methods, our method does not rely on the intricate operation for feature extraction. In this work, only the band-pass filtering and exponential moving standardization are required for processing the raw EEG signals.

1) Band-pass filtering: We employ a third-order Butterworth band-pass filter for capturing most discriminative motor-related band power information. As done in [32], the band-pass filtering with frequency bands of 3–40 Hz is conducted on the raw EEG trials.

2) Exponential moving standardization: For eliminating the undesired signals such as interferences and noise, we employ the electrode-wise exponentially moving standardization to standardize the continuous EEG signals. We select a decay factor a of 0.999 for calculating the mean μ_k and variances σ_k^2 as follows:

$$\mu_k = (1-a)x_k + a\mu_{k-1}, \tag{1}$$

$$\sigma_k^2 = (1-a)(x_k-\mu_k)^2 + a\sigma_{k-1}^2, \tag{2}$$

where x_k represents the input filtered signal at time k. Then the filtered data can be standardized as:

$$\tilde{x}_k = \frac{x_k-\mu_k}{\sqrt{\sigma_k^2}}, \tag{3}$$

where $\tilde{x}_k$ represents the the standardized data.

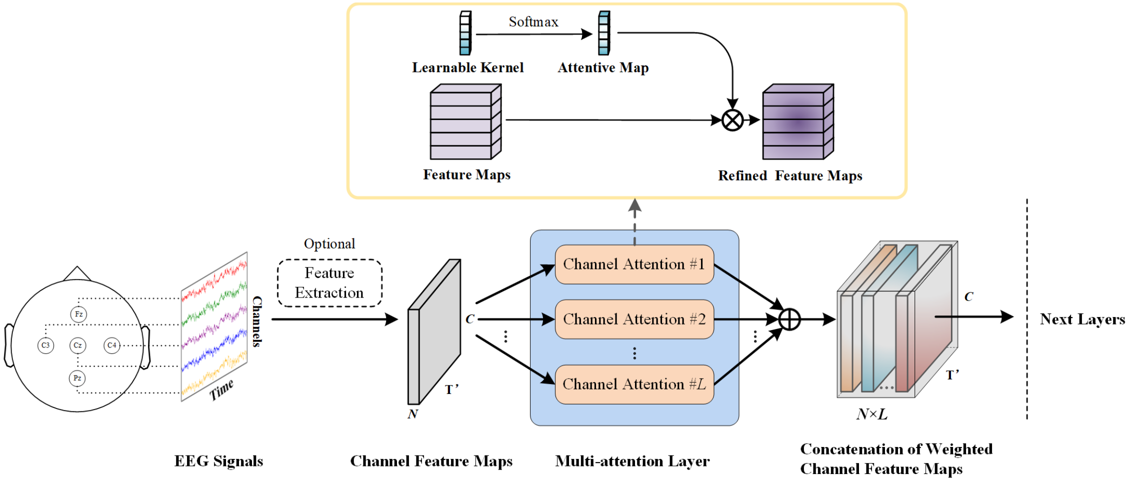


Fig. 2. An illustration of the multi-attention layer in a 5-electrode BCI system. As illustrated, a multi-attention layer with multiple channel attention kernels is utilized to weight the raw EEG signals directly or channel feature maps (when feature extraction is required). Each channel attention kernel employs a learnable tensor and scale its weights with Softmax operation to produce an attentive map. Then the attentive map is applied to the feature maps to weight the channel features. Finally, all the refined feature maps weighted by different kernels are concatenated together and fed to the following layers.

After conducting the above operations on each EEG trial, the derived signals can be considered as cleaner data and preserves the motor imagery information.

III. Methods

In this section, we firstly introduce the basic ideas of the multi-attention layer and its implementation. Then the framework of the proposed MAAN model is presented in detail. Figure 1 illustrates the architecture of the MAAN method, which consists of three basic components: a feature extractor $\mathcal{F}(\cdot)$, a classifier $C(\cdot)$ and a domain discriminator $\mathcal{D}(\cdot)$.

A. Multi-attention Layer

MI signals encode the motor intentions of subjects by modifying the neuronal activity in specific brain regions. According to the previous physiological studies on MI, EEG signals for different brain regions contribute differently to imagination tasks. Concretely, the ERD of hand imagery movements appears over somatosensory areas, while the ERD of feet imagery movements localizes on the central cortex between both hemispheres [11]. Inspired by these neuroscientific findings, we introduce a **Multi-attention Layer** to dynamically estimate the channel-wise contribution in motor intention decoding, aiming at searching the specific spatial patterns relevant to mental tasks. To be concrete, considering a trial of MI signals $x \in \mathbb{R}^{C \times T}$, the multi-attention layer with different kernels is applied to it directly or to its channel feature maps $\mathbf{F} \in \mathbb{R}^{C \times T' \times N}$ as presented in Fig. 2, where T' is the feature dimension, N represents the number of feature sets. Each feature set may be obtained by different feature extraction methods or from different frequency bands. Note that the channel feature maps are equivalent to their corresponding raw signals only when the feature extraction is not required. Next, the features from important brain regions need to be given more attention during the forward process.

For this purpose, the dynamic weight is employed to weight electrodes of different brain regions, which can be denoted by a learnable tensor $\mathbf{W} \in \mathbb{R}^{C \times 1 \times L}$. Here, $\mathbf{W}$ contains L kernels, each kernel corresponds to a specific spatial pattern, denoted by $\mathbf{W}_l = \{w_i\} \in \mathbb{R}^{C \times 1 \times 1}(l = 1, 2, ..., L)$. Note that the proposed multi-attention layer is independent of its input feature maps, like the convolution layer, all the weights are initialized to be 1 in our case. Sequentially, the Softmax operation is applied to each kernel to produce the channel attentive map. Then we weight the input feature maps and output the refined feature maps, as illustrated in the yellow box in Fig. 2. In short, the above attention process can be summerized by the following rule:

$$w_i = \frac{exp(w_i)}{\Sigma_{j=1}^{C} exp(w_j)} \tag{4}$$

and

$$\hat{F}_l = W_l \odot F, \tag{5}$$

where $\odot$ represents the Hadamard product operator. w_i indicates the importance of the i-th channel, which is normalized to [0, 1] using Eq. (4). Finally, all the refined feature maps weighted by multiple kernels are concatenated as a 3D-signal $\hat{F} \in \mathbb{R}^{C \times T \times (N \times L)}$, and fed to the following layers.

As can be seen from the above description, Multi-attention Layer is essentially composed of multiple spatial attention patterns. Each pattern might select some critical channels that are highly relevant to a specific mental task. Thus, it is necessary to deploy multiple patterns in the attention layer, specially in the case of multi-class tasks. Importantly, the parameters of Multi−attention Layer can be dynamically updated by back-propagation algorithm during the training process, which allows learning the appropriate spatial distributions for each subject without requiring any prior knowledge about neuroscience.

TABLE I
MODEL PARAMETERS OF THE MAAN FRAMEWORK

Modules	Layer	Kernel	Stride
Feature extractor	Multi-attention	C×1, 5	1
	Temporal Conv	1×25, 40	1
	Spatial Conv	C×1, 40	1
	Batch Normalization	-	-
	Square Activaction	-	-
	Average Pooling	1×75	15
	Log Activaction	-	-
	Dropout	$p = 0.5$	-
Classifier	Conv-1	1×61, M	1
	Softmax	-	-
Domain discriminator	Conv-2	1×61, 2	1
	Softmax	-	-

B. Feature Extractor

For further extracting the high-level presentations from the EEG signals, we build a multi-attention ConvNet to implement the feature extractor as shown in Fig. 1. It includes the multi-attention layer, the temporal convolutional layer (Temporal Conv), the spatial convolutional layer (Spatial Conv), the batch normalization layer, the Nonlinear Transformation block and the dropout layer. Specially, the size of the convolutional kernel in the temporal layer is 1×25 and in the spatial layer is $C \times 1$ with the stride of 1, which has been shown that such a size of kernel is proper for feature extraction of an EEG time series [32]. In addition, the Nonlinear Transformation block is equipped with squaring nonlinearity layer, average pooling layer and logarithmic activation layer, where the size of the pooling filter is set as 1×75 with the stride of 15. The parameters of the feature extractor are presented in Table 1.

For the multi-channel EEG signals $x \in \mathbb{R}^{C \times T}$ fed into the feature extractor, the multi-attention layer is firstly employed to select the task relevant channels by weighting the input signals. Then the temporal convolutional layer and spatial convolutional layer are applied to learn temporal and spatial representations from the weighted signals. Next, a batch normalization layer is utilized to speed up the convergence speed and improve generalization ability. Finally, the Nonlinear Transformation block is employed for extracting the high-level EEG features. And the operations of dropout are inserted after the Nonlinear Transformation block for further robustness. As a results, we obtain a group of discriminative features from the EEG time series, which are denoted as f. Here, f represents both source domain features f^s and target domain features f^t.

C. Classifier and Domain Discriminator

The classifier (as shown in Fig. 1) is trained to predict the label of high-level EEG features with high certainty. In this work, the classifier is a two-layer network, where the first layer is the convolutional operation (Conv-1) with kernel size 1×61 and M output units, the second layer is the softmax activation function. Here, M is the number of categories for mental tasks.

As can be seen in Fig. 1, the classifier takes feature f as input, then the output $\mathbf{O}$ of Conv-1 is fed into the softmax function to obtain the predicted probability p_i for the k-th category of mental tasks by the following formula:

$$p_k(x) = \frac{exp(o_k)}{\Sigma_{k=1}^{M} exp(o_k)}, \tag{6}$$

where o_k is the k-th element of output $\mathbf{O}$, $k = 1, 2, ..., M$. Consequently, the predicted vector corresponding to feature f can be denoted as $C(f) = (p_1, p_2, ..., p_k)$. Since only the EEG signals from the source domain are labeled, the loss function of the classifier is formulated as:

$$L_c(\theta_f, \theta_c) = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}(C(f_i^s), y_i^s), \tag{7}$$

where θ_f and θ_c represent the parameters in the feature extractor and classifier, respectively. f_i^s is the extracted feature from the source domain, y_i^s is the corresponding ground-truth label, and $\mathcal{L}(\cdot)$ denotes the cross-entropy loss function.

In EEG signal classification, training data and testing data usually come from different distributions, i.e., the training and testing data are drawn from different sessions or subjects. In this case, the classification model trained on training data may perform badly on the testing data. Thus, the domain discriminator is introduced to address this problem. As is shown in Fig. 1, the aim of the discriminator is to distinguish the feature in source domain from target domain. Similarly to the classifier, we adopt a convolutional layer (Conv-2) with kernel size 1×61 and 2 output units as the first layer of the discriminator, and insert softmax operation as the activation function. Considering the given feature sets $F^s = \{f_1^s, f_2^s, ..., f_{n_s}^s\}$ and $F^t = \{f_1^t, f_2^t, ..., f_{n_t}^t\}$ from source and target domains, respectively. Then the loss function of the discriminator can be defined as:

$$L_d(\theta_f, \theta_d) = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}(\mathcal{D}(f_i^s), d_i^s) + \frac{1}{n_t} \sum_{j=1}^{n_t} \mathcal{L}(\mathcal{D}(f_j^t), d_i^t), \tag{8}$$

where θ_d represents the learned parameters in the discriminator, d_i^s and d_j^t denote the source and target domain labels of the input features f_i^s and f_j^t, respectively. During the training, the feature extractor extracts domain-invariant features to reduce the domain discrepancy by maximizing the above loss function of the discriminator.

D. Optimization of Network

During the training stage, the proposed MAAN is jointly optimized by the adversarial learning between classifier and discriminator. To be concrete, the parameters θ_c in the classifier aim to minimize the prediction loss (Eq.(7)) while, simultaneously, the parameters θ_d in the discriminator strive to maximize the domain loss (Eq.(8)). As a result, the parameters θ_f in the feature extractor can be optimized by minimizing the prediction loss to capture the discriminative feature, and by maximizing the domain loss to obtain the domain-invariant feature in such minimax game.

Based on the above analysis, we integrate all components and present the overall loss function of MAAN as follows:

$$L(\theta_f, \theta_c, \theta_d) = L_c - \alpha L_d, \tag{9}$$

where α is the trade-off parameter to balance the classifier and the domain discriminator learning process.

For finding the optimal parameters to minimize the loss function of (9), we can iteratively minimize L_c while maximize L_d. Firstly, the parameters of θ_f and θ_c are updated by minimizing the loss function as follow:

$$(\hat{\theta}_f, \hat{\theta}_c) = \arg\min_{\theta_f, \theta_c} L(\theta_f, \theta_c, \hat{\theta}_d). \tag{10}$$

Subsequently, based on the optimal parameters $\hat{\theta}_f$ and $\hat{\theta}_c$, the parameters of θ_d can be updated by maximizing the following loss function:

$$\hat{\theta}_d = \arg\max_{\theta_d} L(\hat{\theta}_f, \hat{\theta}_c, \theta_d). \tag{11}$$

As a result, the feature extractor is able to learn the discriminative features by minimizing the loss functions L_c. Moreover, in order to maximizing the loss function L_d, the feature extractor generates the domain-invariant features to deceive the discriminator. Consequently, by means of the adversarial learning between classifier and discriminator, we can obtain the discriminative and domain-invariant features for EEG decoding.

Additionally, we insert the *Gradient Reversal Layer* (GRL) the grey part in Fig. 1 [33] before the discriminator to change the maximizing problem into a minimizing problem. During the forward propagation, the GRL acts like an identity transform. However, in the back propagation stage, the GRL reverses the gradient sign in the discriminator by multiplying it with a negative scalar, denoted by -β, here (β=1 in our case). By incorporating the GRL before the discriminator, the parameters of MAAN can be optimized with SGD-based approaches [34]. The optimization process of MAAN is summarized in Algorithm 1.

Algorithm 1 Training procedure of MAAN

```
Input: EEG signals and label from source dataset: $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$,
       EEG signals from target dataset: $D_t = \{(x_i^t)\}_{i=1}^{n_t}$,
       source domain label set: $L_s = \{1\}$,
       target domain label set: $L_t = \{0\}$,
       the learning rate for feature learning and discriminator: $\mu$,
       minibatch size for source and target datasets: $n$,
       adaptation parameter: $\alpha$.
Output: Learned parameters: $\theta_f$, $\theta_c$, $\theta_d$.
Initialize: Initialize model parameters: $\theta_f$, $\theta_c$, $\theta_d$.
while $\theta_f$, $\theta_c$, $\theta_d$ has not converged do
    Source samples $\{(x_i^s, y_i^s)\}_{i=1}^{n}$, a batch from source dataset $D_s$;
    Target samples $\{(x_i^t)\}_{i=1}^{n}$, a batch from target dataset $D_t$;
    Calculate the classifier loss $L_c(\theta_f, \theta_c)$;
    Optimize the parameters of the feature extractor and classifier by:
    $\theta_f \leftarrow \theta_f - \mu \frac{\partial L_c}{\partial \theta_f}$, $\theta_c \leftarrow \theta_c - \mu \frac{\partial L_c}{\partial \theta_c}$;
    Calculate the domain loss $L_d(\theta_f, \theta_d)$;
    Optimize the parameters of the feature extractor and discriminator by:
    $\theta_f \leftarrow \theta_f + \mu \alpha \frac{\partial L_d}{\partial \theta_f}$, $\theta_d \leftarrow \theta_d - \mu \frac{\partial L_d}{\partial \theta_d}$;
end while
```

IV. Experiments

A. Dataset

1) Dataset A (Dataset IIa of BCI competition IV [35]): In this dataset, the EEG signals from nine different subjects are recorded with 22 electrodes during two sessions. Each of them were instructed to perform four classes of MI movements, including movements of the left hand, the right hand, the feet and the tongue. The first session consists of the 288 four-second trials, which is used to be the training set. And the second session consists of the 288 four-second trials, which is used to be the testing set.

2) Dataset B (Dataset IIb of BCI competition IV [36]): This dataset records 3-electrode EEG motor-imagery signals from nine subjects with five sessions of imagined movements of the left or the right hand, the latest 3 sessions include online feedback [36]. The training set consists of the approx. 400 trials of the first 3 sessions, the test set consists of the approx.320 trials of the last two sessions.

3) Additional Dataset (High Gamma dataset [32]): High Gamma dataset (HGD) is a MI dataset created under controlled recording conditions and therefore contains minimum noise. This dataset is recorded using 128 electrodes from 14 healthy subjects, and it consists of a training set of 880 trials and a testing set of 160 trials. In this work, HGD dataset is employed as an additional dataset to investigate whether the performance of our method also hold on other dataset.

B. Experiment Settings

In this work, a variety of state-of-the-art methods are employed for comparing to our framework: FBCSP [17], Sparse Support Matrix Machine (SSMM) [37], ConvNet [32], Channel-wise Convolution with Channel Mixing (C2CM) [38], EEGNet [39], Transfer Component Analysis (TCA) [40] and Joint Distribution Adaptation (JDA) [41] methods . Specifically, FBCSP is a popular baseline for MI signals classification, which utilizes the common spatial pattern feature in different filter bands. SSMM utilizes the low-rank structural information and feature selection to improve the EEG decoding performance. ConvNet is a shallow convolutional neural network tailored to decode band power features. C2CM is a deep convolutional neural network with fine-tuning the model parameters for every subject, such as hidden units, kernel size, and so on. EEGNet is a compact CNN framework designed to decode EEG signals from different BCI paradigms such as P300 visual-evoked potentials, error-related negativity responses, movement-related cortical potentials and sensory motor rhythms. Noted that the EEGNet-10,3 is implemented in this paper because it obtain the best performance in MI classification among the family of EEGNet. TCA is a subspace alignment approach by mapping original high-dimensional representations space into lower-dimensional representations space, while JDA provides an effective framework for transferable representations learning by aligning the joint distributions

TABLE II
The Classification Performance (%) Of different algorithms on dataset A, where stars indicate statistically significant differences between comparison methods and MAAN (Wilcoxon signed-rank test, $p < 0.05$:*, $p < 0.01$:**, $p < 0.001$:***)

Subject	**FBCSP [17]**	**SSMM [37]**	**ConvNet [32]**	**C2CM [38]**	**EEGNet [39]**	**TCA** [40]	**JDA** [41]	**MAAN**
A1	76.00	82.64	76.39	87.50	83.68	75.00	77.78	**86.81**
A2	56.50	60.76	55.21	65.28	63.89	62.15	56.25	**70.49**
A3	81.25	85.76	89.24	90.28	90.97	78.82	81.60	**92.71**
A4	61.00	67.01	74.65	66.67	64.24	63.54	62.50	**85.42**
A5	55.00	58.68	56.94	62.50	59.72	55.56	61.46	**72.22**
A6	45.25	54.51	54.17	45.49	52.08	49.65	49.31	**62.85**
A7	82.75	90.97	92.71	89.58	87.85	74.31	80.90	**93.06**
A8	81.25	81.25	77.08	83.33	82.29	66.32	69.79	**86.46**
A9	70.75	79.51	76.39	79.51	**86.81**	68.40	71.18	83.33
Avg acc (kappa)	67.75(0.570)**	73.45(0.646)**	72.53(0.634)**	74.46(0.660)*	74.61(0.661)*	65.97(0.546)**	67.86(0.571)**	**81.48 (0.753)**

TABLE III
The classification performance (%) of different algorithms on dataset B, where stars indicate statistically significant differences between comparison methods and MAAN (Wilcoxon signed-rank test, $p < 0.05$:*, $p < 0.01$:**, $p < 0.001$:***)

Subject	**FBCSP [17]**	**SSMM [37]**	**ConvNet [32]**	**EEGNet [39]**	**TCA** [40]	**JDA** [41]	**MAAN**
B1	70.00	74.06	76.56	67.50	69.38	68.44	**82.81**
B2	**60.36**	55.00	50.00	60.35	59.29	56.79	**60.36**
B3	**60.94**	55.63	51.56	62.81	55.00	55.00	59.06
B4	**97.50**	94.06	96.88	91.25	93.44	95.00	**97.50**
B5	93.12	86.88	**93.13**	83.44	87.50	87.81	91.88
B6	80.63	82.19	85.31	61.56	75.94	78.44	**86.38**
B7	78.13	76.56	83.75	83.75	78.13	77.81	**84.06**
B8	92.50	92.19	91.56	91.88	90.94	89.38	**93.44**
B9	**86.88**	85.62	85.62	82.5	85.94	82.81	**86.88**
Avg acc (kappa)	80.00(0.600)*	78.00(0.560)**	79.37(0.588)*	76.12(0.682)*	77.28(0.546)**	76.83(0.537)**	**82.54 (0.651)**

TABLE IV
The classification performance (%) for additional dataset, where stars indicate statistically significant differences between comparison methods and MAAN (Wilcoxon signed-rank test, $p < 0.05$:*, $p < 0.01$:**, $p < 0.001$:***)

Method	**FBCSP [17]**	**SSMM [37]**	**ConvNet [32]**	**EEGNet [39]**	**TCA** [40]	**JDA** [41]	**MAAN**
Avg acc (kappa)	90.90(0.879)**	92.50(0.900)**	92.98(0.906)**	85.57(0.808)***	84.81(0.797)***	88.92(0.852)**	**93.97 (0.920)**

of source and target domains, which are widely applied to EEG-based BCIs in recent studies [42]–[45].

Following the competition guideline [46], the training and testing sets are used to train and evaluate the model, respectively. Especially, all examples with labels from training set are used as source domain, and all examples without labels from testing set are used as target domain. Such evaluation protocols is successfully applied to computer vision fields for unsupervised domain adaptation [47], [48]. To make the comparison fair for all methods, we employ the classification accuracy and kappa value as the evaluation metrics. The kappa value is a statistical measurement and can be defined as:

$$k = \frac{acc - p}{1 - p}, \tag{12}$$

where acc and p represent the classification accuracy and the hypothetical probability for the chance agreement, respectively.

Our approach is implemented in Pytorch with an Intel CPU (i7-7700k, 4.2GHz) and an NVIDIA GPU (GTX 1060). For both datasets, all the EEG channels are employed for classification and the three electrooculography (EOG) channels are directly discarded without removing any artifact operation. We train MAAN through back-propagation from scratch. We adopt Adam optimizer with momentum of 0.9, the learning rate of 0.0003 and batchsize = 72 in MAAN during training. And the early-stopping skill [49] is employed for early-terminating the training process if no improvement on the training set is observed in twenty steps to prevent over-fitting.

C. Experimental Results Analysis

We first evaluate different algorithms on the dataset IIa and present classification accuracy on each subject and mean accuracy (kappa) values in Table II. Moreover, the p value of Wilcoxon signed-rank tests [50] is also utilized to check the significant difference of accuracies between the proposals and other state-of-the-art baselines, and the stars shown in Tabel II indicate the levels of significant difference. As is shown, MAAN outperforms all comparison methods and exhibits the statistically significant difference ($p < 0.05$, Wilcoxon signed-rank test), which demonstrates that our architecture is

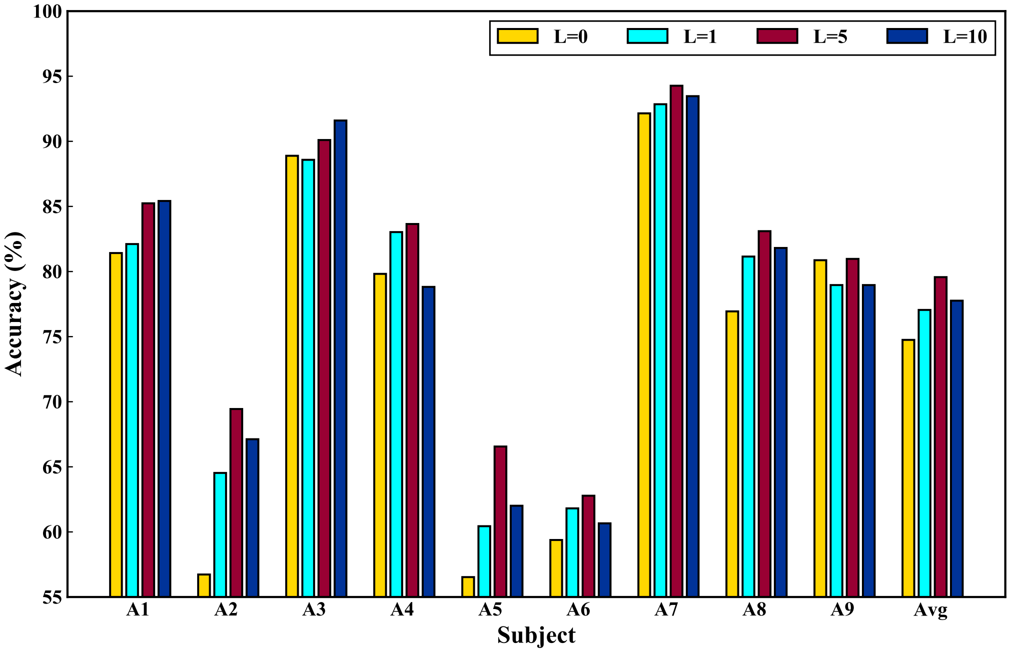


Fig. 3. The performance of MAAN with different multi-attention layer widths on Dataset A.

able to transfer deep presentations across different domains. From the experimental results, we can make the following observations. (1) Deep learning based methods (EEGNet and C2CM) achieve comparable performance and even outperform conventional methods by extracting the handcrafted features, like FBCSP and SSMM; this confirms that the deep neural networks are capable of learning the discriminative features for classification. (2) Comparing with ConvNet, C2CM shows superior performance, implying that fine-tuning the architecture parameters for each subject may improve the classification performance. However, such operation is time consuming and exhausting. (3) The transfer learning methods TCA and JDA shows an inferior classification performance than ours, and the difference between them are statistically significant ($p < 0.01$), suggesting that the domain discrepancy is difficult to be bridged by separately optimizing the feature extraction and classification with minimizing different objective functions. Moreover, such separable optimization procedure is easier to sink into local optima and deteriorate the classification performance of models. In contrast, our deep learning method learns the discriminative representation and classifier in an end-to-end optimization paradigm. Moreover, our framework based on transfer learning technique and multiple attention mechanism is capable of extracting the domain-invariant features and achieving the better performance.

For further verifying the effectiveness of our method, the results on the dataset IIb are reported in Table III. It is noteworthy that our method still outperforms all comparison methods on most subjects, which demonstrates our framework is effective for motor imagery decoding and classification. Looking carefully at Table III, it is surprising that FBCSP yields a better performance than many other methods (SSMM and ConvNet) on Dataset IIb, while it is worse on Dataset IIa. It indicates that handcrafted features may not have enough generalization ability, and the classifier trained with these features is suboptimal. However, the deep learning methods, such as ConvNet and MAAN, show the competitive performance on both datasets, indicating ConvNet is the effective architecture for EEG classification. Especially, the MAAN framework exhibits a powerful EEG decoding ability on MI-based tasks, which contributes to its multi-attention mechanism and domain adaptation. The multiple attentions mechanism target the critical channels automatically to improve the discriminative feature extraction during EEG decoding, while the domain adaptation technique reduces the discrepancy between the signals from different sessions effectively by adversarial learning. All the findings verify the effectiveness and robustness of the proposed framework.

The performance of the additional dataset is shown in Table IV, MAAN obtains the mean accuracy of 93.97% and mean kappa value of 0.920, much better than other deep learning methods, such as SSMM, ConveNet and EEGNet. However, EEGNet only reaches the accuracy of 85.57%, worse than FBCSP with 90.90%.

D. Effectiveness Analysis

In this section, we conduct the following experiments to investigate the effects of parameter settings and to validate the significance of the multiple attentions mechanism and domain adversarial learning.

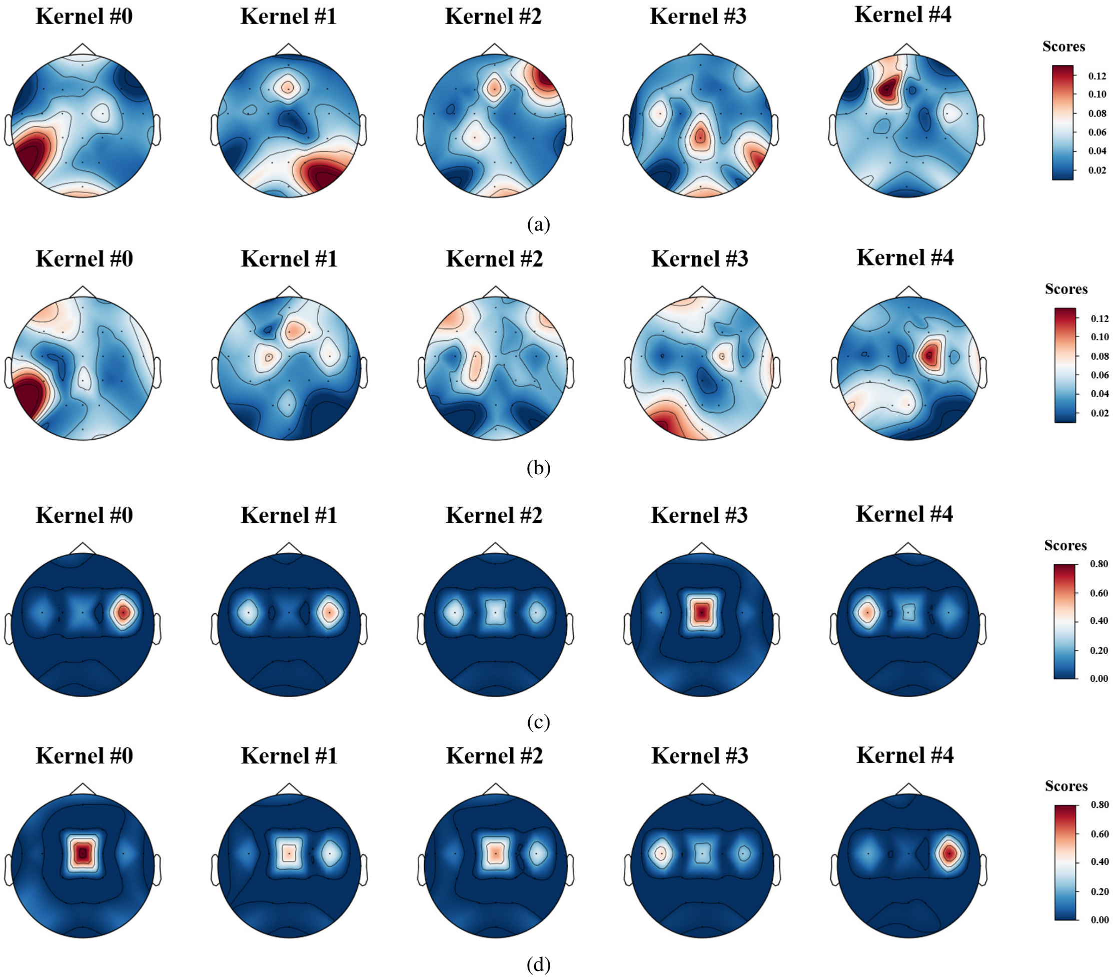


Fig. 4. The visualization of multiple attention kernels learned by multi-attention layer: (a) Subject A6; (b) Subject A9; (c) Subject B6; (d) Subject B9. A large value indicates a major impact on motor imagery classification; in contrast a small value represents minor impact. # represents the digital label of the kernel.

1) Effect of Multi-attention mechanism: To validate the importance of the introduced multiple attentions mechanism, we analyze experimentally the behavior of our model with different multi-attention widths on Dataset IIa. In this experiment, we remove the discriminator module from the proposed framework and set the multi-attention layer width L as {0, 1, 5, 10}. The results presented in Fig. 3 indicate the following conclusions: (1) Comparing the results with L=0, the networks with channel attention mechanism obtain better classification performance in most subjects (e.g., A1, A2, A5, A6, A7, A8). However, some networks without attention mechanism show an higher accuracy than those with attention mechanism. Take subject A3 as an example, the model with L=0 yields a rise of 0.31% in the accuracy compared with L=1, being also worse than those models with L=5 or L=10. Similar phenomenon can be seen in subjects A4 and A9, which demonstrates that it is effective to introduce the channel attention mechanism for improving the EEG decoding performance. (2) Among the networks with attention mechanism, the networks with multiple attention kernels achieve a superior performance than those with single attention kernel, implying that multiple attention kernels capture more informative spatial patterns for classification. (3) However, the performance of MAAN deteriorates when increasing the multi-attention layer width to 10. This result can be found in almost all subjects (except for subjects A1 and A3). The average accuracy indicates that the excessive increase of the multi-attention layer width may lead the network learning to be more noisy, redundant and less informative about the spatial patterns, and cause the overfitting phenomenon. Empirically, we set the width of the multi-attention layer as 5 for our framework to obtain the expected performance.

For further revealing the spatial activation patterns for motor imagery learned by multi-attention layer, the visualization of multiple attention kernels is conducted on two evaluation datasets. For a clear illustration, the parameters of 5 attention kernels are obtained and mapped over the brain surface according to the electrode positions. Figure 4 displays the channel scores estimation of four randomly selected subjects A6, A9, B6 and B9. Actually, similar phenomena also occur for other subjects. In all plots, red regions mark channel relevance for the classification task whereas blue regions mark irrelevant ones. As presented in Figs. 4(a)-(d), the attention kernels only focus on relatively small detailed brain regions, which

TABLE V
THE PERFORMANCE OF MAAN WITH DIFFERENT WEIGHTS OF DOMAIN LOSS ON DATASET A

Subject	***L*=0**			***L*=5**		
	α **= 0**	α **= 0.5**	α **= 1**	α **= 0**	α **= 0.5**	α **= 1**
A1	81.42	86.08	83.42	85.24	**86.81**	84.72
A2	56.73	60.49	54.48	69.44	**70.49**	67.01
A3	88.89	89.72	87.78	90.10	92.71	**93.06**
A4	79.82	82.92	80.14	83.65	**85.42**	84.38
A5	56.53	58.44	66.18	66.56	72.22	**72.57**
A6	59.38	60.87	56.42	62.78	**62.85**	59.38
A7	92.15	92.85	93.68	94.27	93.06	**95.83**
A8	76.94	85.10	84.51	83.10	**86.46**	84.72
A9	80.87	78.92	80.45	80.97	**83.33**	80.90
Avg acc (kappa)	74.75 (0.663)	77.27 (0.697)	76.34 (0.685)	79.57 (0.728)	**81.48 (0.753)**	80.28 (0.737)

is important for EEG feature extraction. For example, the kernel #0, kernel #2, kernel #3 and kernel #4 for subject A6 emphasizes the frontal and central areas (e.g., Fz, FC1, FCz, FC4, C1, C2, C3, C4, CP3, CP1 and CPz) of human brain; the similar activated patterns can be seen in the attention kernels for subject A9, B6 and B9. These brain regions are exactly overlaying the somatosensory cortex and the primary hand cortex, which demonstrates that the primary sensory-motor cortex is activated during MI (hand imagery, feet imagery and tongue imagery). This finding agrees well with the notion confirmed in previous studies [51]–[53]. Moreover, it can be seen that the activation levels on both hemispheres are asymmetric in some attention kernels, such as kernel #0 and kernel #1 for subject A6, kernel #0 and kernel #4 for subject B6, and so on. This phenomenon may be relevant to the ERD/ERS of hand imagery [11]. Interestingly, the multi-attention layer also learns to locate the other EEG channels that might be ignored, such as P1, Pz and POz, implying that these channels could be helpful to discriminative feature extraction. Furthermore, it can be observed that the spatial attention patterns and activation levels cross subjects are different, which correctly explains the intrinsic individual dependency of motor imagery. Therefore, the same channel subsets selection is not suitable for all subjects, it again verifies the effectiveness of the introduced multiple attentions mechanism.

2) *Effect of Domain Adversarial Learning*: We further investigate the effect of domain loss with different weights on Dataset IIa and display the results in Table IV. In this experiment, we consider two different scenarios: MAAN without multi-attention layer (L=0) and MAAN with multi-attention layer (L=5). When L=0, the performance of MAAN increases from 74.75% to 77.27% and 76.34% as α varies from 0 to 0.5 and 1, respectively. The 2.52% and 1.59% improvements of average accuracy demonstrate that the domain adversarial learning reduces the discrepancies between source and target domains effectively. The same conclusion can also be obtained in the case of L=5. Moreover, the results of MAAN with domain loss and multiple attentions exhibits a superior performance than MAAN with only domain loss. Especially, the improvement is even more significant for the subjects with poor signals quality, e.g., subject A2 and A5 yields a rise of 10% and 13.78% in the accuracy, respectively, when α is 0.5. It indicates that transferring the knowledge from the critical brain regions estimated by multi-attention layer is easier and more effective than transferring the information from the whole brain. It is worth noting that the performance of MAAN would deteriorate when the weight value of α increases from 0.5 to 1 in both scenarios, which emphasizes the importance of balancing the classifier and domain discriminator learning process.

3) *$\mathcal{A}$-distance*: To study how the multiple attentions mechanism and domain adversarial learning impact on the distribution discrepancy between sources and targets, the $\mathcal{A}$-distance [54], [55] is employed to measure the distribution discrepancy across domains. According to the previous studies, the $\mathcal{A}$-distance is defined as following:

$$dist_{\mathcal{A}}(\mathcal{D}_s, \mathcal{D}_t) = 2(1 - 2\epsilon), \tag{13}$$

where ϵ represents the test error of a binary classifier (kernel SVM in our case) trained to distinguish source samples from target samples. Figure 5 presents the $dist_{\mathcal{A}}$ with features learned by MAAN with three different settings of multi-attention layer when the weight value of domain loss varies from 0 to 1, including (1) L=0, (2) L=1 and (3) L=5. Note that all the $dist_{\mathcal{A}}$ are averaged on 9 subjects from Dataset IIa. For all the weighted values of domain loss, we observe that $dist_{\mathcal{A}}$ on MAAN-0 features is larger than $dist_{\mathcal{A}}$ on both MAAN-1 and MAAN-5 features, implying that the features learned with multiple attentions mechanism can reduce the domain gap more effectively. The $dist_{\mathcal{A}}$ on MAAN features learned by domain adaptation shows smaller values, demonstrating the effectiveness of the domain adversarial learning. Moreover, for the results in α=0.5 and α=1, the MAAN with multiple attentions exhibits the smaller $dist_{\mathcal{A}}$, which validates the introduction of multiple attentions mechanism that enhances the transferability of features. All the results with α=0.5 show a smaller $dist_{\mathcal{A}}$ demonstrating that a superior classification performance is achieved by adjusting the domain adaptation parameter α to 0.5.

4) *Confusion Matrices*: We exploit the comprehensive analysis of classification results for multi-class motor imagery task by confusion matrixes and the experiment results on test set

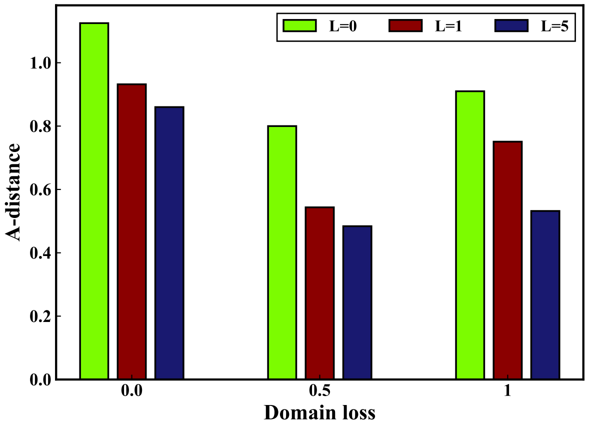
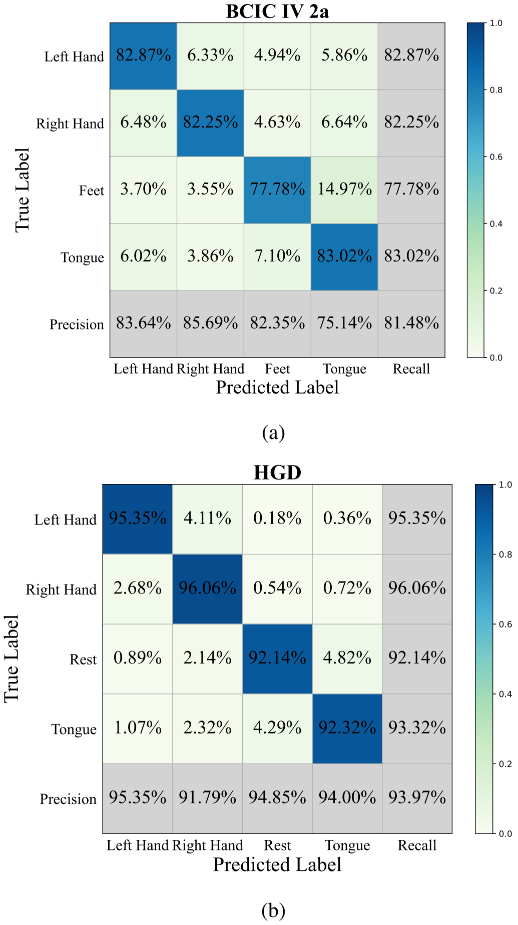


Fig. 5. The $\mathcal{A} - distance$ w.r.t different settings of MAAN on Dataset A.

are given in Fig. 6, where the value in the diagonal line of the confusion matrixes is the correctly predicted samples of the classification task of each motor imagery. For a more comprehensive evaluation of the performance of MANN, we also add recall and precision to the confusion matrix, where the last row of the confusion matrix is the recall for the classification task for each motor imagery and the last column is the precision. From Fig. 6(a), our method could calssify each motor imagery task with high accuracy, and the majority of all mistakes were due to the discriminating between Left Hand/Right Hand and Feet/Tongue. Similarly, Fig. 6(b) showes the same trend on the HGD dataset that the average accuracy of all the four classes is over 92%, which demonstrates the efficacy of MAAN.

5) Feature Visualizations: To verify the transferability of features learned by MAAN, we randomly select subjects A2, A5 and B5 from the two datasets and visualize their network representations as learned by MAAN-0 (without domain adaptation and multi-attention) and MAAN using t-SNE embeddings [56], which are presented in Figs. 7(a)-(c), respectively. Each row corresponds to a different subject; the left plot presents the MAAN-0 features while the right presents the MAAN features. The red dots are features from the previous session (source domain), and the blue triangles from the current session for the same subject (target domain). The visualization results lead to the following observations: (1) With MAAN-0 features, the distributions between the source and the target domains are not aligned very well for both subjects. (2) By contrast, for the feature learned with MAAN, the distributions between the source and the target domains are aligned much better, implying that the domain discrepancies between them are efficiently reduced by transfer learning and multiple attentions mechanism. These observations demonstrate that our method is capable of learning the domain-invariant features, which also explains the superior classification performance of our method.



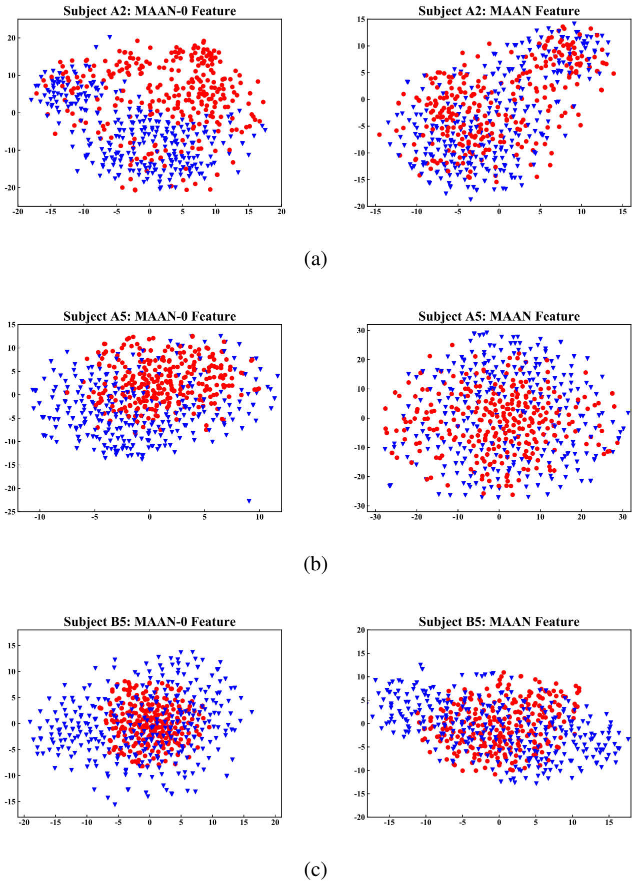
(a)

(b)

Fig. 6. Confusion matrices of MAAN on different datasets. (a) BCIC IV 2a; (b) HGD.

V. Conclusion

In this paper, we propose a novel Multi-Attention Adaptation Network (MAAN) architecture to extract more discriminative and transferable features for EEG signals classification. Firstly, the multi-attention layer is introduced to dynamically estimate the channel-wise contribution during training, which can be easily implemented for end-to-end learning. Then, a multi-attention ConvNet is employed as feature extractor for learning discriminative features from raw EEG signals. Then, the extracted features are fed into the classifier for label prediction. Specially, a domain discriminator is introduced for adversarial domain adaptation on the extracted features, aiming at aligning different distribution across source and target domains. As a result, we can make good use of the source data to learn the discriminative features and to train a robust classifier with better generalization for the target domain. The extensive experiments demonstrate that, compared to other state-of-the-art methods, the proposed method achieves superior performance, contributing to domain-invariant features extraction. Additionally, we visualize the multi-attention layer from the feature extractor, and we confirm that the multiple attentions mechanism can locate the critical channels and different spatial patterns related to MI tasks. It is worth noting that any explicit prior knowledge of the mental task is not



(a)

(b)

(c)

Fig. 7. Feature visualization of three randomly selected subjects by t-SNE. (a) Subject A2; (b) subject A5; (c) subject B5. Red dots, trials from source domain; blue triangles, trials from target domain.

incorporated in the multi-attention layer, it still targets the channels that are important from a viewpoint of physiology whereas irrelevant channels are discarded. Since our method is based on the domain adversarial learning, it is not necessary to collect labelled data for new sessions and it is more convenient to be deployed in EEG-based BCI applications.

However, the current MAAN may still have some limitations. On the one hand, it only leverages the data from previous sessions, the data from other subjects are not considered in the training. On the other hand, we only consider the alignment of marginal distributions between source domain and target domain, though the discrepancy of the distributions for labeling (conditional distribution) between them may not decrease. Our future study will focus on the research about inter-subject transfer and conditional distribution alignment for overcoming the limitations of multi-attention adaptation network.

References

[1] A. E. Hassanien and A. Azar, "Brain-computer interfaces," vol. 74.cham, Switzerland: Springer, 2015.

[2] 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, 2016.

[3] Z. Gao, X. Wang, Y. Yang, C. Mu, Q. Cai, W. Dang, and S. Zuo, "EEG-based spatio-temporal convolutional neural network for driver fatigue evaluation," *IEEE Trans. Neural Netw. Learn Syst.*, vol. 30, no. 9, pp. 2755–2763, 2019.

[4] Y. Yang, Z. Gao, Y. Li, Q. Cai, N. Marwan, and J. Kurths, "A complex network-based broad learning system for detecting driver fatigue from EEG signals," *IEEE Trans. Syst., Man, Cybern., Syst.*, DOI:10.1109/TSMC.2019.2956022, 2019.

[5] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, 2019.

[6] Z. Gao, Y. Li, Y. Yang, N. Dong, X. Yang, and C. Grebogi, "A coincidence filtering-based approach for CNNs in EEG-based recognition," *IEEE Trans. Industr. Inform.*, vol. 16, no. 11, pp. 7159–7167, 2020.

[7] K. J. Olfers and G. P. Band, "Game-based training of flexibility and attention improves task-switch performance: near and far transfer of cognitive training in an EEG study," *Psychol. Res.*, vol. 82, no. 1, pp. 186–202, 2018.

[8] W. He, Y. Zhao, H. Tang, C. Sun, and W. Fu, "A wireless BCI and BMI system for wearable robots," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 7, pp. 936–946, 2016.

[9] Z. Gao, W. Dang, M. Liu, W. Guo, K. Ma, and G. Chen, "Classification of EEG signals on VEP-based BCI systems with broad learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, DOI:10.1109/TSMC.2020.2964684, 2020.

[10] Z. Gao, T. Yuan, X. Zhou, C. Ma, K. Ma, and P. Hui, "A deep learning method for improving the classification accuracy of SSMVEP-based BCI," *IEEE Trans. Circuits Syst. II, Express Briefs*, vol. 67, no. 12, pp. 3447–3451, 2020.

[11] G. Pfurtscheller and F. L. Da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, 1999.

[12] F. Pichiorri, G. Morone, M. Petti, J. Toppi, I. Pisotta, M. Molinari, S. Paolucci, M. Inghilleri, L. Astolfi, F. Cincotti *et al.*, "Brain–computer interface boosts motor imagery practice during stroke recovery," *Ann. Neurol.*, vol. 77, no. 5, pp. 851–865, 2015.

[13] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *J. Neural Eng.*, vol. 15, no. 3, p. 031005, 2018.

[14] H. Ramoser, J. Muller-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, 2000.

[15] B. Chen, Y. Li, J. Dong, N. Lu, and J. Qin, "Common spatial patterns based on the quantized minimum error entropy criterion," *IEEE Trans. Syst., Man, Cybern., Syst.*, pp. 1–12, 2018.

[16] P. Gaur, H. Gupta, A. Chowdhury, K. McCreadie, R. B. Pachori, and H. Wang, "A sliding window common spatial pattern for enhancing motor imagery classification in EEG-BCI," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[17] 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, 2012.

[18] S.-H. Park, D. Lee, and S.-G. Lee, "Filter bank regularized common spatial pattern ensemble for small sample motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 498–505, 2017.

[19] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "A multi-class EEG-based BCI classification using multivariate empirical mode decomposition based filtering and riemannian geometry," *Expert Syst. Appl.*, vol. 95, pp. 201–211, 2018.

[20] P. Gaur, A. Chowdhury, K. McCreadie, R. B. Pachori, and H. Wang, "Logistic regression with tangent space based cross-subject learning for enhancing motor imagery classification," *IEEE Trans. Cogn. Dev. Syst.*, early access, 2021, doi:10.1109/TCDS.2021.3099988.

[21] A. Gupta, R. K. Agrawal, J. S. Kirar, J. Andreu-Perez, W. Ding, C. Lin, and M. Prasad, "On the utility of power spectral techniques with feature selection techniques for effective mental task classification in noninvasive BCI," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 5, pp. 3080–3092, 2019.

[22] B. Pourbabaee, M. J. Roshtkhari, and K. Khorasani, "Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2095–2104, 2018.

[23] C. Yin, S. Zhang, J. Wang, and N. N. Xiong, "Anomaly detection based on convolutional recurrent autoencoder for IoT time se-

ries," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, 2020, doi:10.1109/TSMC.2020.2968516.

[24] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted boltzmann machines," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 566–576, 2016.

[25] 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, 2019.

[26] Z. Gao, W. Dang, X. Wang, X. Hong, L. Hou, K. Ma, and M. Perc, "Complex networks and deep learning for EEG signal analysis," *Cogn. Neurodyn.*, vol. 15, no. 3, pp. 369–388, 2021.

[27] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong, "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition." in *27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 1561–1567.

[28] E. Jeon, W. Ko, and H.-I. Suk, "Domain adaptation with source selection for motor-imagery based BCI," in *7th Int. Winter Conf. Brain-Comput. Interface (BCI)*. IEEE, Feb. 2019, pp. 1–4.

[29] B. Graimann, J. E. Huggins, S. P. Levine, and G. Pfurtscheller, "Visualization of significant ERD/ERS patterns in multichannel EEG and ECoG data," *Clin. Neurophysiol.*, vol. 113, no. 1, pp. 43–47, 2002.

[30] J. K. Feng, J. Jin, I. Daly, J. Zhou, Y. Niu, X. Wang, and A. Cichocki, "An optimized channel selection method based on multifrequency CSP-rank for motor imagery-based BCI system," *Comput. Intell. Neurosci.*, vol. 2019, no. 8068357, pp. 1–10, 2019.

[31] T. N. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf, "Support vector channel selection in BCI," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, 2004.

[32] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[33] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[34] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

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

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

[37] Q. Zheng, F. Zhu, J. Qin, B. Chen, and P.-A. Heng, "Sparse support matrix machine," *Pattern Recognit.*, vol. 76, pp. 715–726, 2018.

[38] 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, 2018.

[39] Lance, B. J., Gordon, S. M., Solon, A. J., Lawhern, V. J., Waytowich, and N. R., "EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, p. 056013 (17pp), 2018.

[40] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, 2011.

[41] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *2013 IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2200–2207.

[42] W. Mu and B.-L. Lu, "Examining four experimental paradigms for EEG-based sleep quality evaluation with domain adaptation," in *2020 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2020, pp. 5913–5916.

[43] X. Jiang, K. Xu, and W. Chen, "Transfer component analysis to reduce individual difference of EEG characteristics for automated seizure detection," in *2019 IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, 2019, pp. 1–4.

[44] W. Hang, W. Feng, R. Du, S. Liang, Y. Chen, Q. Wang, and X. Liu, "Cross-subject EEG signal recognition using deep domain adaptation network," *IEEE Access*, vol. 7, pp. 128 273–128 282, 2019.

[45] B. Wang, W. Li, W. Fan, X. Chen, and D. Wu, "Alzheimer's disease brain network classification using improved transfer feature learning with joint distribution adaptation," in *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2019, pp. 2959–2963.

[46] 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, 2012.

[47] G. Cai, Y. Wang, L. He, and M. Zhou, "Unsupervised domain adaptation with adversarial residual transform networks," *IEEE Trans. Neural Netw. Learn Syst.*, vol. 31, no. 8, pp. 3073–3086, 2020.

[48] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5385–5394.

[49] R. Caruana, S. Lawrence, and C. L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Adv. Neural Inf. Process. Syst.*, 2001, pp. 402–408.

[50] F. Wilcoxon, *Individual Comparisons by Ranking Methods*. New York, NY: Springer New York, 1992, pp. 196–202. [Online]. Available: https://doi.org/10.1007/978-1-4612-4380-9_16

[51] G. Pfurtscheller and C. Neuper, "Motor imagery activates primary sensorimotor area in humans," *Neurosci. Lett.*, vol. 239, no. 2-3, pp. 65–68, 1997.

[52] C. Stippich, H. Ochmann, and K. Sartor, "Somatotopic mapping of the human primary sensorimotor cortex during motor imagery and motor execution by functional magnetic resonance imaging," *Neurosci. Lett.*, vol. 331, no. 1, pp. 50–54, 2002.

[53] V. Morash, O. Bai, S. Furlani, P. Lin, and M. Hallett, "Classifying EEG signals preceding right hand, left hand, tongue, and right foot movements and motor imageries," *Clin. Neurophysiol.*, vol. 119, no. 11, pp. 2570–2578, 2008.

[54] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1-2, pp. 151–175, 2010.

[55] C. Yu, J. Wang, Y. Chen, and M. Huang, "Transfer learning with dynamic adversarial adaptation network," in *2019 IEEE Int. Conf. Data Min. (ICDM)*, 2019, pp. 778–786.

[56] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.

Peiyin Chen received the master's degree in automation with the School of Electrical Engineering and Automation, Wuhan University, Wuhan, China, in 2019.

She is currently pursuing the Ph.D. degree in control science and engineering at the School of Electrical and Information Engineering, Tianjin University, Tianjin, China.

Her research interests include machine learning, deep learning and brain-computer interface.

Zhongke Gao received the M. Sc. and Ph.D. degrees from Tianjin University, China in 2007 and 2010, respectively.

He is currently a Full Professor with the School of Electrical and Information Engineering, Tianjin University, and the Director of the Laboratory of Complex Networks and Intelligent Systems, Tianjin University. He has published over 100 peer-reviewed journal articles.

His current research interests include deep learning, multiphase flows, complex networks, multi-source information fusion, sensor design and brain-computer interface.

Miaomiao Yin Dr. Miaomiao Yin is a Rehabilitation Physician, she obtained her Master degree from Tianjin Medical University in China. Since 2013, she has been working in Huanhu hospital in Tianjin. Currently, she is focusing on basic and clinical research for vestibular rehabilitation.

Jialing Wu Dr. Jialing Wu is a Chief Physician, PhD Supervisor, and he is the Director of the Department of Neurology and Rehabilitation. Dr. Wu obtained his MD degree from Huaxi Medical University and his PhD degree from Tianjin Medical University in China. He went on to his postdoctoral study in ischemic stroke in the Department of Neurology at Emory University. Currently, he is focusing on basic and clinical research for Stroke and Parkinsonism.

Kai Ma received the Ph.D. degree from University of Illinois at Chicago in 2014.

He is currently working as a principal researcher at Tencent. Before joining the current position, he worked for Siemens Medical Solution (US) for more than five years.

Celso Grebogi got his PhD in Physics from the University of Maryland in 1978, Postdoc in Physics and Applied Mathematics at UC Berkeley in 1978-1981.

He is the Sixth Century Chair, and the Founding Director of the Institute for Complex Systems and Mathematical Biology, King's College, University of Aberdeen, UK. He is also an External Scientific Member (Mitglied) of the Max-Planck-Society. He was previously with the University of Sao Paulo as Full Professor of Physics, and, before that, with the University of Maryland as Full Professor of Mathematics.

He has made a major impact for his work in the field of chaotic and complex dynamics. He was awarded the Senior Humboldt Prize and the Thomson-Reuters Citation Laureate. The seminal work on chaos control (OGY) was selected by the American Physical Society as a milestone in the last 50 years. He received multiple Doctor Honoris Causa degrees, Humboldt Senior Prize, Fulbright Fellowship, Toshiba Chair, and various Honorary Professorship awards. He is Fellow of the Royal Society of Edinburgh, The World Academy of Sciences, Academia Europaea, Brazilian Academy of Sciences, American Physical Society, and the UK Institute of Physics.