Deep Representation-Based Domain Adaptation for Nonstationary EEG Classification

He Zhao, *Member, IEEE*, Qingqing Zheng, *Member, IEEE*, Kai Ma, *Member, IEEE*, Huiqi Li, *Senior Member, IEEE*, and Yefeng Zheng, *Senior Member, IEEE*

***Abstract*— In the context of motor imagery, electroencephalography (EEG) data vary from subject to subject such that the performance of a classifier trained on data of multiple subjects from a specific domain typically degrades when applied to a different subject. While collecting enough samples from each subject would address this issue, it is often too time-consuming and impractical. To tackle this problem, we propose a novel end-to-end deep domain adaptation method to improve the classification performance on a single subject (target domain) by taking the useful information from multiple subjects (source domain) into consideration. Especially, the proposed method jointly optimizes three modules, including a feature extractor, a classifier, and a domain discriminator. The feature extractor learns the discriminative latent features by mapping the raw EEG signals into a deep representation space. A center loss is further employed to constrain an invariant feature space and reduce the intrasubject nonstationarity. Furthermore, the domain discriminator matches the feature distribution shift between source and target domains by an adversarial learning strategy. Finally, based on the consistent deep features from both domains, the classifier is able to leverage the information from the source domain and accurately predict the label in the target domain at the test time. To evaluate our method, we have conducted extensive experiments on two real public EEG data sets, data set IIa, and data set IIb of brain–computer interface (BCI) Competition IV. The experimental results validate the efficacy of our method. Therefore, our method is promising to reduce the calibration time for the use of BCI and promote the development of BCI.**

***Index Terms*— Cross subject, deep neural network (DNN), domain adaptation, electroencephalography (EEG), motor imagery (MI).**

I. Introduction

MACHINE learning techniques show great efficiency to extract discriminative information from electroencephalography (EEG) recordings for mental intention recognition and, therefore, play a critical role in EEG-based brain–computer interface (BCI). Providing a nonmuscular channel of communication between the human brain and external devices by translating mental intention into control commands, such BCI systems inspire many promising applications, including communication, movement, and rehabilitation for patients [1], [2], as well as entertainment for healthy people [3].

The deep neural network (DNN), as a subcategory of machine learning, has achieved impressive progress in computer vision [4], [5] and natural language processing [6]. It has been shown that DNN is well suited for end-to-end learning without *a priori* knowledge of the target problem and is able to scale well to a large data set. However, due to the special characteristics of EEG signals, DNN is seldom explored in EEG signal analysis. On the one hand, EEG signals are generally sample-limited and high-dimensional [7]. DNN may suffer from a severe overfitting problem with limited EEG samples since it normally requires a larger amount of training data than other machine learning methods. On the other hand, EEG signals have large intersubject variability, where a subject-independent classifier directly trained on EEG data from multiple subjects often has poor generalization capability on a new subject. Therefore, we cannot simply increase the data size by aggregating the training samples from multiple subjects and feed them into the DNN models.

Domain adaptation [8] was first explored for domain invariant feature learning in computer vision and has been used to overcome this bottleneck of poor performance for every single subject. It is capable to make good use of one specific domain with enough training data, to effectively extract important information or train classifiers adapted to a related but different domain, where only limited labels, even none, can be acquired [9]. In the case of EEG data, we refer to the domain with enough annotated data from multiple subjects as the source domain and the one with limited annotations or none at all from the target subject, as the target domain. Until recently, only a few studies have investigated domain adaptation with DNN in the context of EEG classification. Sakhavi and Guan [10] fine-tuned a convolutional neural network (CNN) pretrained on the source brain signals using target samples with pseudolabels. However, this would result in overfitting when fewer target data are available. To handle limited training data, Li *et al.* [11] proposed a bihemisphere domain adversarial neural network (BiDANN) model for emotion recognition that tackled the marginal distribution shift between training and test data. After that, the classifier trained from labeled source data was applied to target data directly. However, the BiDANN is strongly dependent on the neuroscience findings on emotion

Manuscript received September 29, 2019; revised May 28, 2020; accepted July 11, 2020. This work was supported in part by the Key Area Research and Development Program of Guangdong Province, China, under Grant 2018B010111001, in part by the National Key Research and Development Project under Grant 2018YFC2000702, and in part by the Science and Technology Program of Shenzhen, China, under Grant ZDSYS201802021814180. *(Corresponding author: Qingqing Zheng.)*

He Zhao is with the Beijing Institute of Technology, Beijing 100081, China, and also with Tencent, Shenzhen 518057, China.

Qingqing Zheng, Kai Ma, and Yefeng Zheng are with Tencent, Shenzhen 518057, China (e-mail: aileenzheng@tencent.com).

Huiqi Li is with the Beijing Institute of Technology, Beijing 100081, China.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3010780

and cannot be applied to the motor imagery (MI) task, which is the focus of this article.

To tackle this issue, in this article, we propose a novel end-to-end neural network model, called deep representation-based domain adaptation (DRDA), to deal with EEG MI tasks. Our DRDA model learns the deep feature representation by considering the marginal and conditional distribution discrepancy between source and target domains. This is achieved by jointly optimizing three modules.

1) *Feature Extractor:* It learns the discriminative MI information by mapping the original EEG data into deep features.
2) *Classifier:* It predicts the output labels with the extracted deep features from the feature extractor.
3) *Domain Discriminator:* It is designed to distinguish which domain (source or target) the deep features come from so as to constrain the entire deep feature distribution to be similar across domains.

In addition, to further leverage the label information, the center loss [12] is employed to constrain the invariant feature mapping in the feature extractor. During training, the parameters of feature extractor and classifier are optimized by minimizing the classification loss and center loss together with the adversarial loss provided by the domain discriminator, while the parameters of domain discriminator are updated by maximizing the adversarial loss. This leads to a process of adversarial learning between feature extractor and domain discriminator.

The major contributions of this article can be summarized as follows.

1) To our best of our knowledge, the proposed DRDA model is the first work that explores an end-to-end DNN model with feature space adaptation for MI tasks. By reducing the distribution discrepancy between related but different domains, it is able to leverage source samples to improve the single-subject performance in the target domain, even with fewer labels available.
2) We propose an efficient combined loss function that consists of an adversarial loss to reduce the intersubject discrepancy and a center loss to constrain the intrasubject nonstationarity. Together with the spatial–temporal network, our model can generate effective and nonhandcrafted deep representation of EEG patterns that are discriminative with regard to MI tasks but nondiscriminative with respect to different domains.
3) We extensively evaluate the proposed DRDA model on two real EEG data sets. The experimental results show that DRDA achieves state-of-the-art generalization performance in single-trial EEG-based MI tasks. Therefore, the proposed method has great potential for efficient and robust EEG-based BCIs.

The rest of this article is organized as follows. In Section II, we mainly review the relevant studies on deep learning and domain adaptation used in MI classification. In Section III, we describe the proposed model in detail. The experiments and their results are presented and discussed in Section IV. Finally, Section V concludes this study.

II. Related Works

The conventional common spatial pattern (CSP) method [13] and its various extensions [14], [15] have achieved promising results. The conventional CSP employs a single fixed frequency band to compute the optimal spatial filter such that the ratio of the filtered variance between two categories is maximized (or minimized). Similar to CSP, the filter bank CSP (FBCSP) [16] decomposes the fixed frequency into multiple nonoverlapped subbands and stacks the CSP features in each band. To reduce the computational burden in FBCSP, the discriminative filter band CSP (DFBCSP) [17] employs Fisher's ratio of the spectral power to select the most discriminative frequency bands. However, the extracted matrix-form features are stacked into vectors and fed to a support vector machine (SVM) or a linear discriminant analysis (LDA) classifier, which would inevitably destroy the latent structural information within the raw EEG data. To address this issue, modern matrix-form classifiers have been developed to preserve and leverage the structural correlation by introducing certain constraints on the regression matrix [18]. Zhou and Li [19] proposed a novel model to regularize the rank of logistic regression by the nuclear norm. Luo *et al.* [20] investigated a spectral elastic net regularization and proposed a support matrix machine (SMM) model. Based on SMM, Zheng *et al.* [21] proposed a sparse SMM model (SSMM) to simultaneously consider low-rank structural information and feature selection, which further improved the EEG classification performance.

Recently, DNN methods have been investigated in EEG classification tasks [22]. For instance, Kumar *et al.* [23] employed a multilayer perceptron (MLP) to replace the commonly used classifiers, such as SVM, while keeping the CSP feature extraction mechanism. Yang *et al*. [24] extracted augmented CSP features from varying frequency bands and then fed into CNN for further feature learning and classification. Sakhavi *et al*. [25] proposed a channelwise convolution with channel mixing (C2CM) model to classify the temporal–spatial features based on FBCSP. Bashivan *et al*. [26] converted the EEG time series into spectral topography images by short-time Fourier transform (STFT) and then employed CNN to classify the transformed EEG images. Similarly, Tabar *et al*. [27] used 2-D time–frequency EEG maps generated by STFT as input to a CNN with a stacked autoencoder (SAE) for classification. The aforementioned neural network methods still involve preprocessing, such as specialized feature extraction or STFT for image mapping. In this regard, several studies explored end-to-end network models for EEG feature extraction and classification. For example, Tang *et al*. [28] employed a CNN architecture with several 1-D convolutional layers for raw EEG data. Schirrmeister *et al*. [29] proposed a deep CovNet model using separated temporal and spatial filters and achieved performance as competitive as the widely used FBCSP. Wang *et al.* [30] discussed different models on MI tasks with the input of signals in the frequency domain. They found that CNN achieved the best performance and was proven to be robust for offline analysis. She *et al*. [31] proposed a semisupervised version of the extreme learning machine,

which could utilize both labeled and unlabeled data to increase the classification accuracy.

However, it is difficult to collect enough EEG samples to train the deep models with hundreds of thousands of parameters. A classifier trained on pooled data from multiple subjects generally leads to poor performance since the EEG patterns vary from subject to subject [32]. To address this issue, domain adaptation has been applied to either adapt the features/classifier from the source domain to the target domain or extract common features that are robust for different domains [9]. Based on the shallow version of [29], Dose *et al*. [33] introduced subject-specific adaptation to improve the performance of a single subject. Sakhavi and Guan [10] trained a CNN model on the FBCSP features from multiple subjects and then transferred the model parameters by fine-tuning on the new subject's data. Raza *et al*. [34] presented a covariate shift-detection (CSD) method and retrained the classifier once the convariate shifts were detected. Samek *et al*. [35] constructed an invariant subspace for CSP features by removing the principal nonstationary subspace. Similarly, Song *et al*. [36] developed an adaptive CSP method to classify EEG data from multisubjects, which updated the spatial filters for the target domain during classification. Jeon *et al*. [37] proposed an adaptation approach by using samples of the other subjects. They first selected a source subject whose signal power spectral density was similar to the target one and trained the model with both selected and target subjects. The classifiers for the source and target domains were adaptively trained with a gradient reversal layer unit. A weighted transfer method was proposed in [38], where a classification model was trained on the target data with the cross-entropy loss. Besides, a regularization loss was proposed under the assumption that there was common information across the subjects. The feature distribution similarity between the source and target subjects was employed to weight the regularization of different source subjects in the loss term. He and Wu [39] proposed a data alignment framework in the Euclidean space, where the covariance matrices were aligned between different subjects and could be used as features for classification. Different from those approaches that adapt the spatial filters or the established classifiers from the source domain to the target domain, we seek invariant deep representations between source and target domains with an adversarial learning process.

III. Method

In this section, we first briefly introduce the notations and definitions that are used later in this work. Then, we give an overview of the architecture of our method and illustrate the proposed method in detail.

A domain $\mathcal{D}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_i\}_{i=1}^{N} \in \mathcal{X}$ and N is the sample size. Given a specific domain, a task $\mathcal{T}$ represents a label space $\mathcal{Y}$ and an objective function, which also refers to a conditional probability $P(Y|X)$ from a probabilistic view. In general, $P(Y|X)$ can be learned from the labeled data $\{(x_i, y_i)\}_{i=1}^{N}$ in a supervised manner. In the context of EEG-based MI, we assume that there are two different but related domains, namely, the target domain $\mathcal{D}^t$ and the source domain $\mathcal{D}^s$. Especially, each time, we let $\{(x_i^t, y_i^t)\}_{i=1}^{N_t}$ represent N_t trials of EEG series from a single subject as target data and $\{x_j^s, y_j^s\}_{j=1}^{N_s}$ denote the labeled data of N_s samples from the other subjects, where $x_i^t \sim \mathcal{D}^t, x_j^s \sim \mathcal{D}^s \in \mathbb{R}^{C\times T}$ denote the ith and jth trials of EEG data collected from C electrode channels and T time points in the target and source domains, respectively, and $y_i^t, y_j^s \in \{1, \ldots, \text{cls}\}$ denote the corresponding ith and jth labels, respectively, of cls categories. Then, our goal is to learn a model $\mathcal{M}$ such that the task of $P(Y^t|X^t)$ obtains higher accuracy when given the EEG signals $\hat{x}^t$ from the target subject at test time.

To achieve this goal, this work proposes a novel deep representation-based domain adaptation model, namely, DRDA, to leverage the useful information from both $\mathcal{D}^s$ and $\mathcal{D}^t$ and finally improves the prediction performance in $\mathcal{D}^t$. The main idea is to learn a shared feature space across domains via adversarial learning such that the important information within the raw EEG series can be easily extracted and decoded without complicated preprocessing. To implement this idea, DRDA is formulated as a combination of a feature extractor, a domain discriminator, and a classifier, resulting in the overall architecture in Fig. 1(a). First, the feature extractor involving temporal–spatial convolutional operators is engaged to learn the deep feature representations for EEG series. Intuitively, such features are drawn from different distributions for source and target domains since the marginal distribution discrepancy of the raw data is tremendous between different domains. In this way, it is quite difficult to simultaneously leverage the useful information from source features and target features to train the classifier. To address this issue, the domain discriminator is employed to narrow the feature's distribution discrepancy between the source domain and the target domain in the deep representation space. To further learn the discriminative features, we also introduce the center loss to reduce the nonstationarity in the target domain, which pushes away the features from different categories while pulling closer the features belonging to the same class. Then, the deep features learned from both domains can be used to train the classifier. At the test time, as shown in Fig. 1(b), the combination of the feature extractor and the classifier forms an end-to-end deep learning model, which directly predicts the MI label from the EEG input. In what follows, we will introduce the whole process of our approach, including the preprocessing step, the end-to-end network architecture, and the loss functions.

A. Preprocessing

Different from conventional methods [13], [15], [16], [25], which depend on complicated preprocessing procedures for feature extraction, our method learns the deep representation features with the feature extractor module. Given the raw EEG series, only the bandpass filtering and standardization are required to process the data before fed into our model. Overall, the minimal preprocessing without any handcrafted feature design provides a concise and flexible pipeline.

1) Bandpass Filtering: Based on [25], a third-order Butter-worth bandpass filter is employed to filter out the

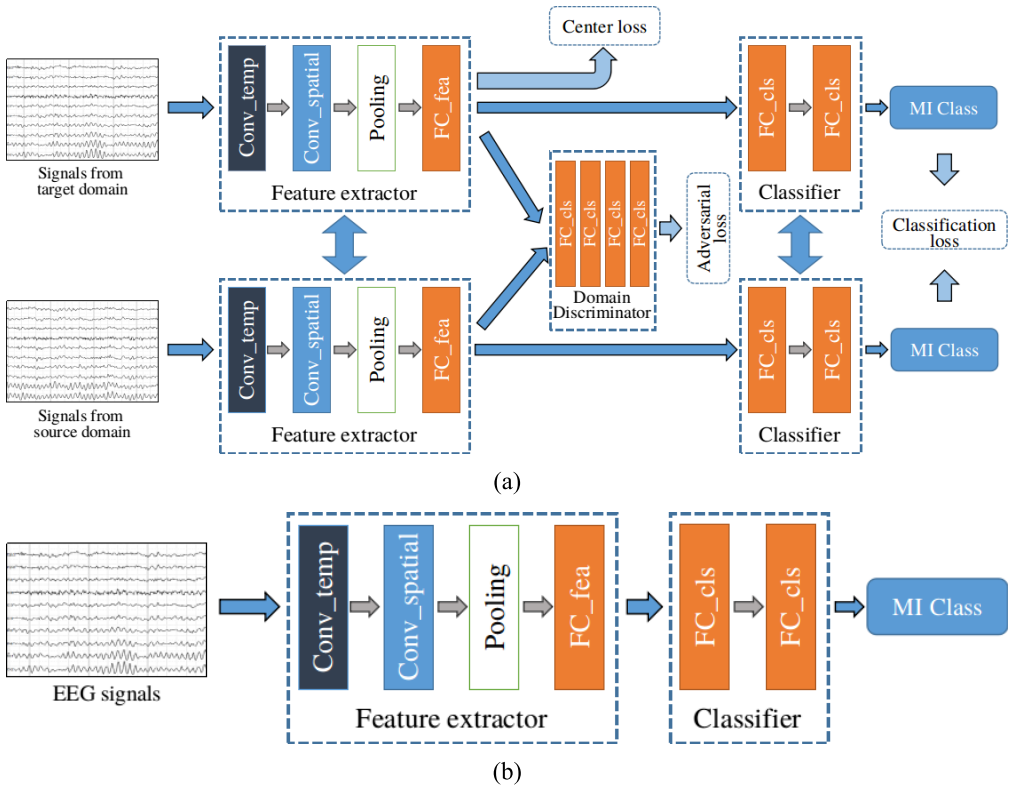

Fig. 1. Architecture of the proposed DRDA method. (a) Training phase of our model. (b) Test phase of our model.

artifacts and disentangle sensorimotor rhythms. In this work, the raw EEG signals are bandpass filtered to [4, 38] Hz.

2) Exponential Moving Standardization: To reduce the non-stationarity and fluctuations, the electrodewise exponential moving standardization is performed to standardize the bandpass filtered data. The standardization can be formulated as

$$x'_k = \frac{x_k - \mu_k}{\sqrt{\sigma_k^2}} \tag{1}$$

where x'_k and x_k denote the standardized and input signal at time k, respectively. μ_k and σ_k^2 are the exponential moving mean and variance values calculated as

$$\mu_k = (1-\alpha)x_k + \alpha\mu_{k-1} \tag{2}$$

$$\sigma_k^2 = (1-\alpha)(x_k - \mu_k)^2 + \alpha\sigma_{k-1}^2 \tag{3}$$

where α denotes the decay factor and is set to be 0.999. In the beginning, we set μ_0 and σ_0^2 to be the mean value and variance of each electrode in a trial. For each time point k, the signal x_k is standardized by data from the past, which eliminates the occasional motion of the signal and preserves the trend of each trial. Note that the preprocessing operations are trial-independent, so it is applicable for an online BCI.

B. Network Architecture

As shown in Fig. 1(a), at the training step, the proposed model consists of three components, including a feature extractor, a classifier, and a domain discriminator. We first adopt a Siamese-like structure to take advantage of data from both domains. Namely, the feature extractor and classifier share their weights for both $\mathcal{D}^s$ and $\mathcal{D}^t$. However, the features learned are different between source and target domains due to the marginal distribution shift caused by fatigue or distraction of subjects during signal collection. To reduce the marginal distribution shift, a domain discriminator is employed to constrain the learned features from $\mathcal{D}^s$ and $\mathcal{D}^t$ to be as close as possible. Therefore, different from the conventional methods, the proposed method is able to make good use of the pooled data from other subjects (source domain) to improve the single-subject (target domain) performance.

1) Feature Extractor: The input 2-D EEG data $x \in \mathbb{R}^{C\times T}$ are often associated with a large size in the temporal dimension and contains full of temporal–spatial structural information [7]. Thus, we cannot simply treat the EEG series as a special kind of natural images and directly perform 2-D convolutional operations to explore the temporal information. Inspired by the success of FBCSP [40], which used bandpass and CSP spatial filtering steps for temporal–spatial transformations, we split the 2-D convolution into two 1-D convolution operations for EEG classification. Especially, the first two layers of the feature extractor are composed of two 1-D convolutional layers to learn the temporal and spatial feature representations, respectively. For temporal convolution, a kernel size of 25 is applied to deal with the continuous and extensive series. It is much larger than the one that is usually used (3×3) in image classification task [5], allowing for a larger range of temporal transformations in this layer. For spatial convolution, a kernel size of the same number of electrode channels C is engaged to fuse the spatial information from all input electrodes. This operation fuses the spatial information from

TABLE I

MODEL PARAMETERS OF THE FEATURE EXTRACTOR AND CLASSIFIER, WHERE C IS THE NUMBER OF ELECTRODES AND CLS INDICATES THE NUMBER OF CLASSES

Modules	Layers	Kernel
Feature Extractor	Temporal Conv	1×25, 30
	Spatial Conv	$C \times 1$, 30
	Average Pooling	1×75, stride15
	FC	64
Classifier	FC	64
	FC	cls

different electrodes to the features of a single electrode. Afterward, an average pooling layer of size 75 is applied to avoid overfitting and learn the invariant features. Finally, a fully connected (FC) layer is connected to generate a deep representation. The details of the feature extractor are shown in Table I.

2) Classifier: To decode the deep representations from the feature extractor, we propose a simple yet efficient classifier with FC layers. Especially, the classifier consists of two FC layers, which is followed by a softmax function to transform the network predictions into class labels. The architecture details of the classifier are presented in Table I. Note that the conditional distributions from source and target domains may also be mismatched due to the subject's mental state, namely, $P(Y^s|X^s) \neq P(Y^t|X^t)$. Therefore, the classification performance may be deteriorated if the classifier trained only with the source domain is applied to target data directly. To address this issue, we consider the conditional distribution inconsistency and utilize the labeled features from both domains to train a robust classifier. If the true labels of target data are absent, we can turn to a pseudolabel strategy [10], which first estimates the pseudolabels of target data under the assumption of $P(Y^s|X^s) \approx P(Y^t|X^t)$ and then updates the parameters of the classifier with the annotated source data and target data with pseudolabels.

3) Domain Discriminator: This module is essential in our method, which offers the ability to leverage data from other subjects in the target domain. Inspired by the generative adversarial network (GAN) [41], we design the adversarial learning process with the feature extractor and domain discriminator components. During the learning process, the discriminator distinguishes the features learned from which domain; at the same time, the feature extractor learns to map the EEG input from both domains into a latent common space. Finally, the feature extractor is able to fool the discriminator such that the domain discriminator fails to distinguish the origin of the extracted features. In this way, the marginal distribution discrepancy of the latent features is alleviated so that the learned features from both domains can be used to train the same classifier. For implementation, the discriminator is trained on a binary domain label set $\mathcal{Z} = \{0, 1\}$, in which the domain label is 1 for target data and 0 for the source samples. As shown in Fig. 1, the domain discriminator consists of four FC layers with size of 64, 32, 16, and 1, respectively. The activation function is set to *relu* for the first three layers and *sigmoid* for the last layer to output a probability for binary prediction.

C. Loss Function

The proposed method jointly optimizes the feature extractor, classifier, and domain discriminator. To be specific, we iteratively train these three modules in two alternative steps and update the parameters according to the chain rule in deep learning methods. At each iteration, we first update the parameters of the domain discriminator, fix the feature extractor and classifier, and then fix the domain discriminator and update the parameters of both the feature extractor and classifier. During training, several loss functions are adopted to measure the difference between network predictions and the given ground truth, which is demonstrated as follows.

For the domain discriminator, it regards the features from source data as fake samples while those from target data as real samples. The domain adaptation operations are carried on the deep representation space generated by the feature extractor. Based on LS-GAN [42], we adopt a least squared adversarial loss to update the discriminator, whose formulation is given as follows:

$$\min_D \mathcal{L}_D = \frac{1}{2}\mathbb{E}_{x^s \sim \mathcal{D}^s}|D(F(x^s)) - a|^2 + \frac{1}{2}\mathbb{E}_{x^t \sim \mathcal{D}^t}|D(F(x^t)) - b|^2 \tag{4}$$

where $F(\cdot)$ denotes the feature extractor and $D(\cdot)$ is the domain discriminator. $F(x^s)$ and $F(x^t)$ denote the latent features of $\mathcal{D}^s$ and $\mathcal{D}^t$, respectively, from the feature extractor. In addition, we choose $a = 0$ to state that $F(x^s)$ comes from the source domain and $b = 1$ for $F(x^t)$ from the target domain.

To reduce the distribution discrepancy, the feature extractor plays a minimax game with the domain discriminator. Based on the feedback of the discriminator, the feature extractor aims to narrow the feature distribution gap between the source and the target domains and, finally, fool the discriminator with a similar feature distribution across domains. Therefore, the adversarial loss for updating the feature extractor F is formulated as

$$\min_F \mathcal{L}_{\text{adv}} = \frac{1}{2}\mathbb{E}_{x^s \sim \mathcal{D}^s}|D(F(x^s)) - c|^2 \tag{5}$$

where $c = 1$. With this adversarial loss, it would lead to a balanced status that the feature extractor finally generates the latent features that are indistinguishable from any domain. In this way, the source data have a similar feature distribution to the target domain such that it can be safely used to train a robust classifier for the target domain.

For the classification, we employ a cross entropy loss to minimize the difference between predictions of model $\mathcal{M}$ and corresponding ground truth with

$$\min \mathcal{L}_{\text{cls}} = -\mathbb{E}_{x \sim \mathcal{D}^s \cup \mathcal{D}^t} \sum_{k=1}^{\text{cls}} \mathbb{1}_{(y==k)} \log(\mathcal{M}(x)) \tag{6}$$

where $\mathbb{1}$ is the indicator function, which is set to be 1 if the condition $y == k$ is satisfied or 0 if not.

It is also well-known that the MI EEG series may vary from session to session even for the same subject, which makes the EEG classification a challenging problem. To tackle this problem, we further study the latent features and employ

a center loss [12] to minimize the intraclass variation and maximize the interclass distance at the same time with

$$\min \mathcal{L}_{ct} = \frac{1}{2}\mathbb{E}_{x^t \sim \mathcal{D}^t} \| F(x^t) - \mathbf{c}_{y^t} \|_2^2 \tag{7}$$

where $\mathbf{c}_{y^t} \in \mathbb{R}^d$ denotes the y^tth class center of deep representation of the target data. Note that the center loss is only applied to the target features since we are more concerned about the classification of the target domain.

Finally, the objective function to jointly update the feature extractor and the classifier is formulated as

$$\min \mathcal{L} = \omega_{\text{cls}}\mathcal{L}_{\text{cls}} + \omega_{\text{adv}}\mathcal{L}_{\text{adv}} + \omega_{ct}\mathcal{L}_{ct} \tag{8}$$

where ω_{cls}, ω_{adv}, and ω_{ct} are weights for the classification loss, adversarial loss, and center loss, respectively.

IV. Experiments

In this section, we extensively validate the proposed method on EEG-based MI classification in the context of BCI. First, we introduce two public EEG data sets used in the following experiments, Data set IIa and Data set IIb of BCI Competition IV for multiclass and binary classification, respectively. Second, since we are not aware of any existing end-to-end deep learning method with domain adaptation for the MI task, we compare our method with several state-of-the-art methods that demonstrate satisfactory performance on these two data sets. Finally, we perform an ablation study of the proposed method with respect to the model hyperparameters, namely, different types of adversarial loss as well as the corresponding weights of different loss terms.

A. Data Description

1) Data Set IIa of BCI Competition IV: The data set[1] [40] contains 22-channel EEG signals from nine subjects (refer to A01–A09). The sampling rate of the signals is 250 Hz. The data were collected on four different MI tasks, including the left hand, right hand, tongue, and both feet. For each subject, two sessions of data were collected with 288 trials (72 trials per MI task) for each session. Here, the MI data in the first session were used for training, and those in the second session were used for the test. Note that the temporal segment of [2, 6] *second* is considered in our experiments.

2) Data Set IIb of BCI Competition IV: The data set[2] [45] records three bipolar-channel EEG signals from nine subjects, namely, B01–B09 involving left hand and right hand MI activities. The sampling rate is 250 Hz. For each subject, five sessions were collected. The data in the first three sessions were used for training and the rest were used for tests. There are about 400 trials and 320 trials in the training and test sets, respectively. We use the temporal segment of [3, 7] s in our experiments.

[1]http://www.bbci.de/competition/iv/#dataset2a

[2]http://www.bbci.de/competition/iv/#dataset2b

B. Experiment Settings

Our approach is implemented with the TensorFlow library in Python with an Intel Core I7 CPU and a Tesla P40 GPU. For these two data sets, all the EEG channels are utilized for classification, and the three electrooculography (EOG) channels are directly discarded without any artifact removing operation. We train our DNNs with Adam optimizer [46], which has a learning rate of 0.0002. The loss weights ω_{cls}, ω_{adv}, and ω_{ct} in (8) are set as 1, 1, and 0.5, respectively, in the experiments for both data sets IIa and IIb of BCI Competition IV. The network parameters are updated by a minibatch with a size of 64 in each training iteration. We early terminate the training if no improvement in the training set is observed in ten iterations to avoid overfitting. Note that the training and test sets are split according to the competition guideline [40], making the comparison fair for all methods.

Since there is no end-to-end deep learning method with domain adaptation for the MI task, to demonstrate the advantage of our method, we compare our method with the following state-of-the-art methods, including the winner algorithm for both data sets (FBCSP [40]), matrix-form EEG classifiers (SMM [20], SSMM [21]), two deep-learning models (MI-CNN [33], and ConvNet [29]), as well as two domain adaptation methods (CCSP [44] and SSCSP [35]). We also compare with the methods that have achieved competitive performance on either of the two data sets, including C2CM [25], BO [43], HSS-ELM [31] for data set IIa, and TLCSD [34] for data set IIb. For a fair comparison, we either list the results of the compared methods from previous publications or fine-tune all the parameters and present their best performance. For evaluation metrics, we employ the classification accuracy as well as kappa value (κ), which takes account of the accuracy occurring by chance and is denoted as

$$\kappa = \frac{\text{acc} - p_0}{1 - p_0} \tag{9}$$

where acc denotes the classification accuracy and p_0 is the accuracy of random guess.

C. Experimental Results Analysis

We first evaluate different algorithms on data set IIa of BCI Competition IV and present the classification accuracy on each subject and mean accuracy (kappa) values in Table II. For a clear illustration, the highest accuracy or kappa values are highlighted in boldface. The results show that the proposed method has superior performance in both mean accuracy and kappa values. The deep learning methods, including ConvNet, C2CM, and ours, outperform the traditional methods, such as FBCSP. It indicates that the DNNs are able to learn the discriminative features for classification. On the contrary, both traditional feature extraction methods with domain adaptation (CCSP and SSCSP), have inferior classification performance. This may result from the strong prior assumptions of the data. For example, CCSP assumes that the spatial filter of different subjects should be as close as possible and estimates the covariance matrix by shrinking it toward the mean of other subjects. Similarly, SSCSP investigates the shared global

TABLE II
CLASSIFICATION ACCURACY (IN PERCENTAGE %) OF DIFFERENT ALGORITHMS ON DATA SET IIA OF BCI COMPETITION IV

	Subject									Average acc (kappa)
	A01	A02	A03	A04	A05	A06	A07	A08	A09	
MI-CNN [33]	73.26	28.82	89.58	68.06	26.39	28.82	75.35	78.82	77.08	60.69 (0.4758)
HSS-ELM [31]	81.14	49.86	78.02	63.33	44.03	49.44	81.11	81.49	81.38	67.76 (0.5701)
FBCSP [16]	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75 (0.5700)
SMM [20]	81.94	59.38	81.60	62.85	59.03	49.36	86.11	77.78	78.47	70.75 (0.6100)
SSMM [21]	82.64	60.76	85.76	67.01	58.68	54.51	90.97	81.25	79.51	73.45 (0.6460)
C2CM [25]	**87.50**	**65.28**	**90.28**	66.67	**62.50**	45.49	89.58	83.33	79.51	74.46 (0.6595)
ConvNet [29]	76.39	55.21	89.24	74.65	56.94	54.17	**92.71**	77.08	76.39	72.53 (0.6338)
BO [43]	82.12	44.86	86.60	66.28	48.72	53.30	72.64	82.33	76.35	68.13 (0.5751)
CCSP [44]	84.72	52.78	80.90	59.38	54.51	49.31	88.54	71.88	56.60	66.51 (0.5535)
SSCSP [35]	76.74	58.68	81.25	57.64	38.54	48.26	76.39	79.17	78.82	66.17 (0.5489)
Ours	83.19	55.14	87.43	**75.28**	62.29	**57.15**	86.18	**83.61**	**82.00**	**74.75** (**0.6633**)

TABLE III
CLASSIFICATION ACCURACY (IN PERCENTAGE %) OF DIFFERENT ALGORITHMS ON DATA SET IIB OF BCI COMPETITION IV

	Subject									Average acc (kappa)
	B01	B02	B03	B04	B05	B06	B07	B08	B09	
MI-CNN [33]	75.31	57.50	56.56	96.88	92.19	83.44	84.06	92.81	86.25	80.56 (0.6111)
FBCSP [16]	70.00	60.36	60.94	**97.50**	93.12	80.63	78.13	92.50	86.88	80.00 (0.6000)
SMM [20]	67.81	51.79	53.44	93.31	82.81	74.69	72.19	82.50	75.62	72.70 (0.4540)
SSMM [21]	74.06	55.00	55.63	94.06	86.88	82.19	76.56	92.19	85.62	78.00 (0.5600)
ConvNet [29]	76.56	50.00	51.56	96.88	93.13	85.31	83.75	91.56	85.62	79.37 (0.5875)
TLCSD [34]	70.31	50.63	50.81	93.75	63.75	74.06	61.88	83.13	77.19	69.72 (0.3944)
CCSP [44]	63.75	56.79	50.00	93.44	65.63	81.25	72.81	87.81	82.81	72.70 (0.4540)
SSCSP [35]	65.00	56.79	54.06	95.63	74.69	79.06	80.00	87.81	82.81	75.09 (0.5018)
Ours	**81.37**	**62.86**	**63.63**	95.94	**93.56**	**88.19**	**85.00**	**95.25**	**90.00**	**83.98** (**0.6796**)

subspace and removes the principal nonstationary subspace common to most subjects before CSP computation. These assumptions are merely held in real applications since EEG data are nonstationary and changed from subject to subject. In addition, almost all compared methods, including FBCSP, SMM, C2CM, BO, CCSP, and SSCSP, separately optimize the feature extraction and classification by minimizing different objective functions. The extracted features may not be optimized for the subsequent classification. Therefore, our method, which not only learns the discriminative features and classifiers in an end-to-end optimization paradigm but also leverages information from other domains with an adversarial learning scheme, can achieve the best performance. Note that though C2CM also achieves promising performance, it fine-tunes the deep architecture parameters for each subject, such as kernel size and hidden nodes, while ours holds the same model architecture and parameters, making it practical for real online BCI applications.

We further test our method on data set IIb of BCI Competition IV. The results of classification accuracy on each subject with an average score of accuracy (kappa) are reported in Table III.[3] As is shown, our method obtains the highest classification accuracy rates on almost all the subjects and also achieves the highest average accuracy/kappa values compared with other methods. It indicates that CNN with domain adaptation in an end-to-end paradigm is effective for EEG classification. The domain adaption realized by adversarial learning is able to pull the data distribution from different subjects close to each other without any strong assumptions. Therefore, the important information of EEG data from the source domain (other subjects) can be exploited when training a discriminative classifier for the target domain (a specific subject). We also notice that, taking FBCSP as a baseline, the techniques with handcrafted features, such as SMM and SSMM, outperform FBCSP on data set IIa but are beaten on data set IIb. It shows that the handcrafted features extracted with expert knowledge may not have enough generalization ability. It again validates the efficacy and robustness of our method.

[3]Since there is no released code for C2CM yet and it is difficult to reproduce, we do not compare our method with C2CM on this data set.

D. Ablation Study

We conduct the following experiments to study the parameter settings and demonstrate the significance of the adversarial loss and the center loss used in the proposed method. In addition, we also evaluate the pseudolabel strategy when the labels in the target domain are absent. Note that to reduce the randomness, we conduct all the experiments in five times and record the mean accuracy and mean kappa values.

1) Adversarial Loss: Table IV shows the performance of our approach on data set IIa of BCI Competition IV with different adversarial loss settings. When the weight of adversarial loss ω_{adv} is set to be 0, our method degenerates to an end-to-end DNN without domain adaptation module, namely, the domain discriminator in Fig. 1 is excluded from the model. We first only feed in the target data from $\mathcal{D}^t$ to tailor the feature extractor and classifier to the target subject. As is shown, the overall performance (mean accuracy: 61.18%) is inferior compared with those (e.g., FBCSP mean accuracy:

TABLE IV

EEG CLASSIFICATION PERFORMANCE OF MODELS TRAINED WITH DIFFERENT ADVERSARIAL LOSS FUNCTIONS AND WEIGHTS ω_{ADV} ON DATA SET IIA OF BCI COMPETITION IV WITH $\omega_{\text{CLS}} = 1$ AND $\omega_{ct} = 0$

		Subject									Average acc (kappa)
		A01	A02	A03	A04	A05	A06	A07	A08	A09	
w/o adversarial loss	$\mathcal{D}^t$	73.26	26.67	89.58	65.77	31.53	28.61	77.92	81.25	76.04	61.18 (0.4824)
	$\mathcal{D}^t + \mathcal{D}^s$	76.74	26.46	81.81	55.76	41.18	41.94	64.23	75.70	72.43	59.58 (0.4611)
	Fine-tuning	80.00	50.62	86.74	67.50	49.86	52.85	85.42	80.21	81.80	70.56 (0.6074)
Vanilla GAN loss	ω_{adv} = 0.01	80.97	51.18	88.26	71.87	58.54	54.31	82.22	81.25	81.18	72.20 (0.6293)
	ω_{adv} = 0.1	80.97	55.07	88.19	70.97	56.39	54.10	84.03	82.92	82.08	72.75 (0.6366)
	ω_{adv} = 1	79.10	54.24	88.96	74.24	59.17	55.00	85.83	82.99	81.46	**73.44 (0.6459)**
	ω_{adv} = 10	77.85	53.13	87.22	73.89	61.53	55.28	85.04	82.22	82.57	73.19 (0.6425)
LS-GAN loss	ω_{adv} = 0.01	80.62	53.82	89.03	71.32	55.28	52.92	83.54	81.67	81.73	72.21 (0.6295)
	ω_{adv} = 0.1	79.86	53.47	89.24	73.26	60.37	51.80	83.96	82.50	81.74	72.91 (0.6388)
	ω_{adv} = 1	82.36	53.40	88.82	75.49	58.75	55.90	85.49	82.08	82.71	**73.89 (0.6518)**
	ω_{adv} = 10	75.76	55.00	89.51	75.69	58.61	55.55	86.80	83.06	81.12	73.46 (0.6461)

TABLE V

PERFORMANCE OF OUR METHOD WITH DIFFERENT WEIGHTS OF CENTER LOSS ω_{ct} ON DATA SET IIA OF BCI COMPETITION IV WITH $\omega_{\text{CLS}} = 1$, $\omega_{\text{ADV}} = 0$ FOR *w/o DA*, AND $\omega_{\text{ADV}} = 1$ FOR *DA*

		Subject									Average acc (kappa)
		A01	A02	A03	A04	A05	A06	A07	A08	A09	
w/o DA	ω_{ct} = 0	73.26	26.67	89.58	65.77	31.53	28.61	77.92	81.25	76.04	61.18 (0.4824)
	ω_{ct} = 0.5	74.86	37.36	90.21	68.96	35.35	32.64	80.39	80.49	77.08	**64.14 (0.5219)**
	ω_{ct} = 1	76.53	29.65	89.72	67.85	37.22	33.05	83.33	81.04	77.29	63.97 (0.5196)
DA	ω_{ct} = 0	82.36	53.40	88.82	75.49	58.75	55.90	85.49	82.08	82.71	73.88 (0.6518)
	ω_{ct} = 0.5	83.19	55.14	87.43	75.28	62.30	57.15	86.18	83.61	82.50	**74.75 (0.6634)**
	ω_{ct} = 1	83.40	55.42	88.54	75.56	62.22	56.18	86.88	82.01	81.39	74.62 (0.6616)

67.75%) listed in Table II. Interestingly, it is observed that the performance suffers from serious degradation for those subjects whose EEG signals have a poor signal-to-noise ratio (SNR) (e.g. A02, A05, and A06), while, for the rest subjects (A01, A03, A04, and A07–A09), the average accuracy of our method has a competitive advantage against those of the other state-of-the-art methods (77.30% versus FBCSP: 75.50%). It illustrates that for the subject with high SNR EEG signals, our deep model can achieve satisfactory results with little training data, while, for those with EEG data of poor quality, more data may be needed to train a reliable classifier. To augment the data, we are motivated to train a subject-independent classifier with all data from $\mathcal{D}^s$ and $\mathcal{D}^t$ and report the classification accuracy as "$\mathcal{D}^s + \mathcal{D}^t$" in Table IV. Though it improves the performance of A04 and A05 obviously, the mean accuracy drops to 59.58%. Thus, it shows that indiscriminately gathering samples from multiple subjects is inefficient due to the distribution shift of different subjects. Moreover, we also fine-tune the deep model pretrained on the source data using the target data for each subject and report the results in Table IV. Though the fine-tuning method significantly outperforms those in both the abovementioned scenarios, there is still an obvious gap between it and our method with an adversarial loss. It, in turn, validates that the proposed method with the domain discriminator is able to narrow the distribution discrepancy and leverage the useful information from source data to improve the classification performance.

We also investigate the influence of two adversarial losses, namely, vanilla GAN loss [41] versus LS-GAN loss [42], and show the results of different settings of ω_{adv} in Table IV. It is observed that for the vanilla loss, with the increase in ω_{adv}, the accuracy rates also increase, demonstrating that taking into account the EEG data from multiple subjects in an appropriate way is helpful. When ω_{adv} is larger than the optimal value and continues to increase, the classification accuracy decreases. This is because when ω_{adv} is too large, the latent features from both domains tend to be the same, and some additional information from the source domain would be discarded. Similar trends also occur in the case of the LS-GAN loss. In addition, the overall performance of the proposed method with the LS-GAN loss is consistently better than that with the vanilla GAN loss. Therefore, we select the LS-GAN loss and set ω_{adv} to be 1.

2) Center Loss: We finally investigate the influence of the center loss with different weights on data set IIa of BCI Competition IV and display the performance in Table V. Two different scenarios are considered where "w/o DA" stands for the models trained on data $\mathcal{D}^t$ and "DA" represents the models trained on data $\mathcal{D}^t + \mathcal{D}^s$ with domain adaptation. When $\omega_{ct} = 0$, the center loss is inactive, and our method does not consider the intrasubject nonstationarity. It is observed that when $\omega_{ct} \neq 0$, the classification accuracy increases for both scenarios, especially for the model "w/o DA", improving from 0.6118 to 0.6414. The improvement is much more significant for the subjects with signals of low quality, e.g., A02 yields a rise of 11% in the accuracy. This observation suggests that the center loss is able to handle the intraclass variations caused by low SNR or nonstationary signals within a single subject, thus making the features learned more discriminative. For a clear illustration, we further employ the t-SNE method [47] and visualize the feature distribution of two randomly selected

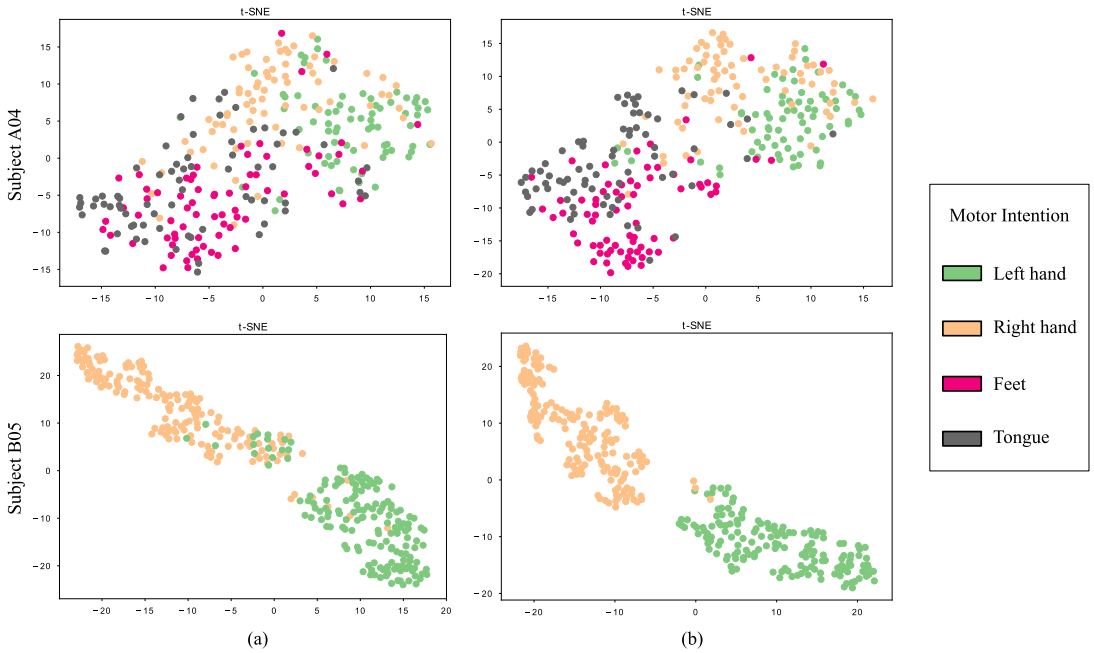

Fig. 2. Visualization of feature distribution by t-SNE [47]. (a) Without the center loss ($\omega_{ct} = 0$). (b) With the center loss ($\omega_{ct} = 0.5$).

TABLE VI
CLASSIFICATION ACCURACY (IN PERCENTAGE %) OF MODEL TRAINED WITH THE PSEUDOLABEL STRATEGY ON DATA SET IIA OF BCI COMPETITION IV

	Subject									Average acc (kappa)
	A01	A02	A03	A04	A05	A06	A07	A08	A09	
Pseudo-label strategy	69.52	26.74	82.50	41.25	27.64	32.36	49.59	76.53	70.14	**52.92 (0.3723)**
Unsupervised DA	68.06	26.06	74.73	39.41	30.90	37.85	43.68	75.00	64.93	51.19 (0.3492)
SourceCNN	67.57	26.53	71.32	45.56	31.04	35.29	40.28	70.77	63.56	50.26 (0.3368)

subjects A04 and B05 from these two data sets, respectively, in Fig. 2. In fact, similar phenomena also occur in other subjects. It also reveals that if without the center loss, the latent features are more scattered and of large entropy. On the contrary, with the help of the center loss, the features are more discriminative, with smaller intraclass distances and larger interclass boundaries. Thus, it leads to better classification performance. Empirically, the center loss weight ω_{ct} is set to 0.5 for our final model.

3) Pseudolabel Strategy: When the labels of target data are absent, we propose a pseudolabel strategy to predict the target labels for domain adaptation. If without the estimated target labels, the center loss would be deactivated, and our method would degrade to an unsupervised domain adaptation method. If the target data are also absent, the domain adaptation would be dysfunctional, and our method would further degrade to the one trained with only the annotated source data (SourceCNN). Thus, to evaluate the pseudolabel strategy in our method, we compare the performance of the pseudolabel strategy, unsupervised domain adaptation, and SourceCNN. Table VI displays the performance of these three models. The results show that our pseudolabel strategy achieves the highest classification accuracy, while SourceCNN obtains the lowest one. It indicates that the classifiers trained directly on the source data have poor generation capability on the target domain. On the contrary, our pseudolabel strategy is capable to improve the target domain classification by leveraging additional information from the source data.

V. CONCLUSION

In this article, we proposed a deep end-to-end domain adaptation method to handle the EEG-based MI classification task, which improves the performance of the subject-dependent classifier by leveraging the useful information from the source domain. To alleviate the distribution discrepancy between the source and target domains, a domain discriminator is engaged to narrow the covariate shift with an adversarial learning strategy. In this way, the features generated from the source domain have a similar distribution to those from the target domain. In addition, a center loss is introduced to learn invariant features, which reduces the intrasubject nonstationarity by minimizing the feature distance of the same class and maximizing the boundary of different categories. Therefore, we can make good use of the source data to learn the discriminative features and train a robust classifier with better generalization for the target domain. We have conducted extensive experiments, and the results show that the proposed approach is effective to identify motor intention from EEG signals and outperforms state-of-the-art methods. Moreover, it also shows empirically that the proposed adversarial loss and the center loss are able to significantly reduce the intersubject

and intrasubject nonstationarity, which can be extended to other BCI applications.

REFERENCES

[1] N. Birbaumer and L. G. Cohen, "Brain–computer interfaces: Communication and restoration of movement in paralysis," *J. Physiol.*, vol. 579, no. 3, pp. 621–636, Mar. 2007.

[2] K. K. Ang *et al.*, "A large clinical study on the ability of stroke patients to use an EEG-based motor imagery brain-computer interface," *Clin. EEG Neurosci.*, vol. 42, no. 4, pp. 253–258, Oct. 2011.

[3] K. J. F. Olfers and G. P. H. 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, Jan. 2018.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2414–2423.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[6] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[7] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 4, no. 2, p. R1, 2007.

[8] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[9] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, Apr. 2019.

[10] S. Sakhavi and C. Guan, "Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI," in *Proc. 8th Int. IEEE/EMBS Conf. Neural Eng.*, May 2017, pp. 588–591.

[11] 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 *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1561–1567.

[12] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 499–515.

[13] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541–1548, Sep. 2005.

[14] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 11, pp. 2274–2281, Nov. 2006.

[15] H. Zhang, H. Yang, and C. Guan, "Bayesian learning for spatial filtering in an EEG-based brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1049–1060, Jul. 2013.

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

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

[18] M. Dyrholm, C. Christoforou, and L. C. Parra, "Bilinear discriminant component analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1097–1111, May 2007.

[19] H. Zhou and L. Li, "Regularized matrix regression," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 76, no. 2, pp. 463–483, Mar. 2014.

[20] L. Luo, Y. Xie, Z. Zhang, and W.-J. Li, "Support matrix machines," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 938–947.

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

[22] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, Jun. 2019, Art. no. 031001.

[23] S. Kumar, A. Sharma, K. Mamun, and T. Tsunoda, "A deep learning approach for motor imagery EEG signal classification," in *Proc. 3rd Asia–Pacific World Congr. Comput. Sci. Eng. (APWC CSE)*, Dec. 2016, pp. 34–39.

[24] H. Yang, S. Sakhavi, K. Keng Ang, and C. Guan, "On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 2620–2623.

[25] 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.

[26] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*. [Online]. Available: http://arxiv.org/abs/1511.06448

[27] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, Feb. 2017, Art. no. 016003.

[28] Z. Tang, C. Li, and S. Sun, "Single-trial EEG classification of motor imagery using deep convolutional neural networks," *Optik*, vol. 130, pp. 11–18, Feb. 2017.

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

[30] Z. Wang, L. Cao, Z. Zhang, X. Gong, Y. Sun, and H. Wang, "Short time Fourier transformation and deep neural networks for motor imagery brain computer interface recognition," *Concurrency Comput. Pract. Exp.*, vol. 30, no. 23, p. e4413, Dec. 2018.

[31] Q. She, B. Hu, Z. Luo, T. Nguyen, and Y. Zhang, "A hierarchical semi-supervised extreme learning machine method for EEG recognition," *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 147–157, Jan. 2019.

[32] X. Chai *et al.*, "A fast, efficient domain adaptation technique for cross-domain electroencephalography(EEG)-based emotion recognition," *Sensors*, vol. 17, no. 5, p. 1014, May 2017.

[33] H. Dose, J. S. Müller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Expert Syst. Appl.*, vol. 114, pp. 532–542, Dec. 2018.

[34] H. Raza, H. Cecotti, Y. Li, and G. Prasad, "Adaptive learning with covariate shift-detection for motor imagery-based brain–computer interface," *Soft Comput.*, vol. 20, no. 8, pp. 3085–3096, Aug. 2016.

[35] W. Samek, F. C. Meinecke, and K.-R. Müller, "Transferring subspaces between subjects in brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 8, pp. 2289–2298, Aug. 2013.

[36] X. Song, S.-C. Yoon, and V. Perera, "Adaptive common spatial pattern for single-trial EEG classification in multisubject BCI," in *Proc. Int. IEEE/EMBS Conf. Neural Eng.*, Nov. 2013, pp. 411–414.

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

[38] A. M. Azab, L. Mihaylova, K. Keng 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.

[39] 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.

[40] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, p. 39, Mar. 2012.

[41] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[42] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2794–2802.

[43] H. Bashashati, R. K. Ward, and A. Bashashati, "User-customized brain computer interfaces using Bayesian optimization," *J. Neural Eng.*, vol. 13, no. 2, Apr. 2016, Art. no. 026001.

[44] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *IEEE Signal Process. Lett.*, vol. 16, no. 8, pp. 683–686, Aug. 2009.

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

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[47] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

He Zhao (Member, IEEE) received the B.E. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 2020 and 2014, respectively.

He was a Research Intern with Tencent, Shenzhen, China, in 2019. His research interests include medical image processing, deep learning, and computer vision.

Qingqing Zheng (Member, IEEE) received the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2018.

She is currently a Senior Researcher with Tencent, Shenzhen, China. Her research interests include machine learning theory, computer vision, and human–computer interaction.

Kai Ma (Member, IEEE) received the Ph.D. degree from the University of Illinois at Chicago, Chicago, IL, USA, in 2014.

He was with Siemens Medical Solution, Princeton, NJ, USA, for more than five years. He is currently a Principal Researcher with the Jarvis Lab, Tencent, Shenzhen, China. His research interests include medical image analysis, deep learning, computer vision, and brain–computer interface.

Huiqi Li (Senior Member, IEEE) received the Ph.D. degree from Nanyang Technological University, Singapore, in 2003.

She is currently a Professor with the Beijing Institute of Technology, Beijing, China. Her research interests include image processing and computer-aided diagnosis.

Yefeng Zheng (Senior Member, IEEE) received the B.E. and M.E. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1998 and 2001, respectively, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2005.

He was a Principal Key Expert with Siemens Healthineers, Princeton, NJ, USA, working on medical image analysis. He is currently the Director of the Jarvis Lab, Tencent, Shenzhen, China. His research interests include medical image analysis, computer vision, and deep learning.