Dynamic Joint Domain Adaptation Network for Motor Imagery Classification

Xiaolin Hong, Qingqing Zheng, *Member, IEEE*, Luyan Liu, *Member, IEEE*, Peiyin Chen, Kai Ma, *Member, IEEE*, Zhongke Gao, *Senior Member, IEEE*, and Yefeng Zheng, *Senior Member, IEEE*

***Abstract*— Electroencephalogram (EEG) has been widely used in brain computer interface (BCI) due to its convenience and reliability. The EEG-based BCI applications are majorly limited by the time-consuming calibration procedure for discriminative feature representation and classification. Existing EEG classification methods either heavily depend on the handcrafted features or require adequate annotated samples at each session for calibration. To address these issues, we propose a novel dynamic joint domain adaptation network based on adversarial learning strategy to learn domain-invariant feature representation, and thus improve EEG classification performance in the target domain by leveraging useful information from the source session. Specifically, we explore the global discriminator to align the marginal distribution across domains, and the local discriminator to reduce the conditional distribution discrepancy between sub-domains via conditioning on deep representation as well as the predicted labels from the classifier. In addition, we further investigate a dynamic adversarial factor to adaptively estimate the relative importance of alignment between the marginal and conditional distributions. To evaluate the efficacy of our method, extensive experiments are conducted on two public EEG datasets, namely, Datasets IIa and IIb of BCI Competition IV. The experimental results demonstrate that the proposed method achieves superior performance compared with the state-of-the-art methods.**

***Index Terms*— Deep neural network (DNN), domain adaptation, adversarial learning, electroencephalogram (EEG), motor imagery (MI), brain-computer interface (BCI).**

I. Introduction

BRAIN computer interface (BCI) systems provide a novel communication pathway for users to manipulate external electrical devices by directly decoding their neuronal activities into specific commands [1]–[5]. To capture such neuronal activities, electroencephalogram (EEG) measures voltage fluctuations on the scalp from multiple electrodes with a fixed spatial arrangement [6]. Due to the great advantages of non-invasiveness and high temporal resolution, EEG has been widely employed in various BCI applications, including military affairs, emotion recognition, driver fatigue detection and epileptic seizure detection [7]–[11]. Since the classification of EEG signals plays a key role in BCI systems, it is highly demanded to develop a reliable and effective EEG decoding scheme to promote the BCI applications.

Recently, many machine learning methods have been investigated for EEG signal decoding. Such methods usually extract discriminative patterns as the first step and then train a classifier, such as linear discriminant analysis (LDA) [12] and support vector machine (SVM) [13] to identify the user's intention. However, these traditional methods depend heavily on handcrafted features that are designed beforehand by human experts, such as frequency band power [14] and common spatial pattern (CSP) [15], [16], which may not cope well with non-stationary EEG signals. Inspired by the excellent performance of deep learning in computer vision, convolutional neural networks (ConvNets) have also been proposed to learn domain agnostic features and non-linear classifiers for EEG decoding. For example, Zheng and Lu [17] explored a probabilistic deep learning algorithm based on deep belief networks for EEG emotion classification. Kumar *et al.* [18] developed a ConvNet framework to classify the CSP features extracted from motor imagery EEG trials. Schirrmeister *et al.* [19] proposed an end-to-end shallow ConvNet architecture for motor imagery recognition. While all aforementioned methods have achieved impressive performance, they generally assumed that test data had the same or similar generation process/distribution as the training set. Yet, in many BCI applications, it is often not the case, since different human mental states and equipment noises may result in large cross-session and cross-subject variance in EEG data. Such shift in data distribution would greatly degrades the performance of well-trained models in test phases.

To tackle these distribution variations, domain adaptation has been applied to BCIs for distribution calibration, where sufficient knowledge annotated EEG signals from previous sessions or subjects (*i.e.*, *source* domain) are transferred to boost the model's performance on the unlabeled data from a new session or subject (*i.e.*, *target* domain) [20]. It aims to

Manuscript received October 8, 2020; revised February 4, 2021; accepted February 5, 2021. Date of publication February 15, 2021; date of current version March 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61873181 and Grant 61922062. *(Corresponding author: Zhongke Gao.)*

Xiaolin Hong and Peiyin Chen are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China, and also with Tencent, Shenzhen 518057, China (e-mail: aaronxlhong@tencent.com; pychen@tju.edu.cn).

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

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

Digital Object Identifier 10.1109/TNSRE.2021.3059166

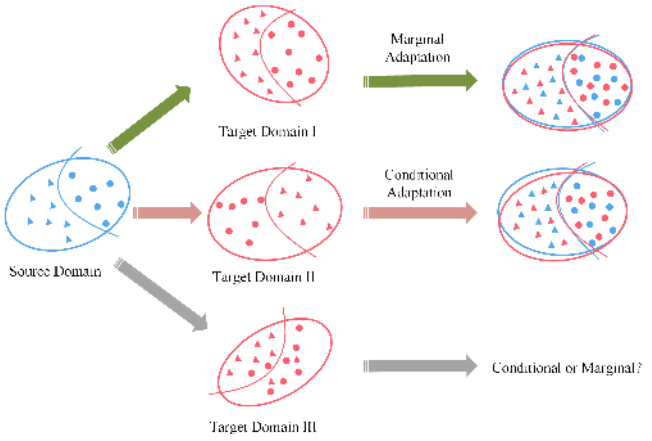

Fig. 1. Triangles and dots denote EEG data belonging to different categories. The EEG data between the source and target domains may be different in both marginal and conditional distributions.

adapt the feature representation or classifier models to reduce the distribution discrepancy between the source and target domains. Traditional methods perform adaptation by either re-weighting source data to achieve a similar distribution as the target one, or discovering an explicit transformation that aligns the feature representation in both domains [21], [22]. For instance, He and Wu [23] aligned EEG trials from different subjects before classification under the assumption that the mean Euclidean covariance matrix should be an identity matrix for all subjects after alignment. While, the latest studies have indicated that domain adaptation combined with deep networks are capable of learning more transferable representation for EEG classification [24].

Recently, inspired by generative adversarial network (GAN), adversarial learning has been embedded in domain adaptation networks in a two-player game mode. In the training process, the deep classification model learns domain-invariant representations with respect to the shift between different domains, while the domain discriminator is used to contest with the classification model, and distinguish which domain the features come from. For instance, Li *et al.* [25] proposed a bi-hemisphere domain adaptation network (BiDANN) for EEG emotion recognition. In BiDANN, a domain discriminator is integrated in a ConvNet framework to reduce the marginal distribution discrepancy between different subjects. In this way, EEG features from different subjects are aligned, and classifiers subsequently trained on aligned EEG features from source subjects can improve the prediction performance for target subjects. However, such domain-adversarial networks only consider the marginal distribution difference but ignore the complex multi-modal structures within EEG data. It may still fail to guarantee that two domains have sufficiently similar distributions between sub-domains even if the marginal distributions are completely aligned. For example, in Fig. 1, labeled source and unlabeled target domains may be different in both marginal (global) and conditional (local) distributions. When marginal distributions of both domains are dissimilar (source domain → target domain I), the alignment of marginal distributions are supposed to be given more attention. When marginal distributions of both domains are very similar (source domain → target domain II), the conditional distributions may be inconsistent between sub-domains. In addition, the alignment of global and local distributions across domains usually contributes differently to the adaptation (source domain → target domain III). While existing domain adaptation methods seldom quantitatively evaluate the relative importance between the global and local distributions.

To address the above issues, in this paper, we propose an unsupervised domain adaptation method, referred to dynamic joint domain adaptation network (DJDAN), for cross-session motor imagery classification. Our DJDAN model learns the domain-invariant feature representation by considering both the marginal and conditional distribution discrepancies between different domains with end-to-end adversarial learning. This is achieved by jointly optimizing four modules, namely, feature extractor, classifier, global and local discriminators. Firstly, the feature extractor based on ConvNet [19] is employed to learn deep feature representation for discriminative motor imagery information. The extracted deep features are subsequently fed into the classifier to predict the output labels. Then, the global domain discriminator is designed to distinguish in a global perspective which domain (source/target) the deep features come from so as to reduce the marginal distribution shift across domains. Similarly, the local domain discriminator is leveraged to constrain conditional distribution inconsistency across sub-domains. However, it is difficult to condition the local domain discriminator on discriminative information since the label information for target domain is unavailable. In this regard, we leverage the discriminative information embedded in the classifier predictions to assist adversarial adaptation for unlabeled target data. Namely, the local discriminator is conditioned on the uncertainty of classifier predictions. In addition, we further introduce a *dynamic adversarial factor* to adaptively evaluate the relative importance of the marginal and conditional distribution alignment during training. When two domains are very different, adaptation pays more attention to the global discriminator for marginal distribution shift. When global distributions are close, the local discriminator is given more attention to align conditional distributions across sub-domains. This leads to a flexible adaptation of global and local adversarial learning between feature extractor and domain discriminators.

The major contributions of this paper are summarized as follows.

- We propose a novel dynamic adversarial network, namely, DJDAN, to learn domain-invariant feature representation for motor imagery task. It is general and does not require to explicitly learn a transformation for feature extractor and classification.
- Our method performs domain adaptation not only by simultaneously considering both the marginal and conditional distribution discrepancies, but also dynamically estimating their relative importance during training.
- We analyze the efficacy of our method and theoretically guarantee a generalization bound on the target error.
- We extensively evaluate the proposed DJDAN model on two public motor imagery datasets (Dataset IIa and Dataset IIb of BCI Competition IV). The experimental

results show that our DJDAN model achieves the state-of-the-art performance.

The remainder of this paper is organized as follows. Section II reviews the related studies on domain adaptation used in EEG signal classification. Section III presents our unsupervised domain adaptation method in details. The experiments and results are presented and discussed in Section IV. Finally, Section V concludes the paper.

II. Related Works

Extracting discriminative features from EEG signals is the key for single-trial EEG classification. Traditional methods such as power spectral density, differential entropy and differential asymmetry have achieved promising results [26]–[29]. Among them, common spatial pattern (CSP) [15], [16] is one of the most popular algorithm to detect discriminative movement-related patterns for motor imagery tasks. It employs a single frequency band to compute the optimal spatial filter such that the ratio of filtered variance between different categories is maximized. Inspired by CSP, filter bank CSP (FBCSP) [30] further decomposes the fixed frequency used in CSP into multiple non-overlapped sub-bands, and stacks the filtered signals in each band as the discriminative features. Moreover, some novel CSP-based algorithms based on feature selection and channel selection methods are proposed to extract effective features [31], [32]. Recently, several deep learning architectures [19], [33]–[35] have been exploited to learn deep representation and classifier for EEG signals in an end-to-end manner. A review of deep learning analysis of EEG signals see [36]. However, these deep methods usually require sufficient annotated data to train the networks with thousands of parameters. Moreover, all aforementioned methods assume that both training and test data are generated from the same or similar distribution. It is often not the case since that different mental states or complex equipment noises may result in distribution shift between the training and the test data. Typically, a classifier trained on the features derived from previous sessions generally suffers performance degradation when tested on those from a new session.

Domain adaptation is a practical and promising technique that leverages prior knowledge learned from the relevant source domain to boost the performance on the target domain, which is helpful to reduce the calibration time and reliance on annotated data required for EEG classification. For example, inspired by maximum mean discrepancies (MMD) [37], He and Wu [23] aligned EEG trials from different subjects in the Euclidean space under the assumption that mean covariance matrices should be an identity matrix, and thus improved the classification performance for new subjects. Azab *et al.* [21] proposed the S-wLTL framework to leverage useful information from similar subjects for training a logistic regression classifier. In S-wLTL, source data are assigned different weights according to the Kullback-Leibler divergence of subject-specific CSP features between the source and target domains. Similarly, Jeon *et al.* [24] investigated another similarity estimation to select source data by using power spectral density of EEG signals in the resting state. Then, both the target data and selected source data are fed into a domain adversarial network. During training, the discrepancy of deep representation from different domains would be reduced by the adversarial learning between the domain discriminator and feature extractor. Tang and Zhang [38] integrated a conditional domain discriminator into a convolutional neural network to learn commonly shared intra-subject EEG features. However, marginal and conditional distribution discrepancies usually contribute differently to the adaptation. These works focus on aligning either the marginal distributions or the conditional distributions, which may fail to simultaneously account for the global and local domain mismatch across domains especially when EEG data are of complex multi-modal structures.

To address these issues, we are motivated to exploit an adversarial neural network to simultaneously consider marginal and conditional distribution adaptation. Inspired by [39], we also utilize a dynamic adversarial factor to adaptively measure the relative importance of marginal and conditional distribution alignment. With this factor, the network is able to dynamically adjust the domain alignment preference during training.

III. Method

A. Notations and Problem Definition

The EEG data collected in a session is defined as $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$, where n represents the number of EEG trials. $\mathbf{x}_i \in \mathbb{R}^{E\times T}$ denotes an EEG trial with E electrodes and T sampling points and $\mathbf{y}_i \in \mathbb{R}^{C}$ is the corresponding label of C categories. In unsupervised domain adaptation, we are given a source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ of n_s annotated EEG trials from previous sessions and a target domain $\mathcal{D}_t = \{(\mathbf{x}_j^t)\}_{j=1}^{n_t}$ of n_t unlabeled trials from a target session. Both $\mathcal{D}_s$ and $\mathcal{D}_t$ share the same feature space and label space, *i.e.*, $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{E\times T}$ and $\mathbf{y}_i, \mathbf{y}_j \in \mathbb{R}^{C}$. The source and target domains are sampled from different joint distributions $P_s(\mathbf{x}^s, \mathbf{y}^s)$ and $P_t(\mathbf{x}^t, \mathbf{y}^t)$, respectively. When $P_s(\mathbf{x}^s, \mathbf{y}^s) \neq P_t(\mathbf{x}^t, \mathbf{y}^t)$, it may be insufficient to adapt only the marginal distribution of the feature representation [21], [23]. For instance, in the real scenario of multi-class classification, even if the feature distributions are similar, there is no guarantee that multi-modal distributions are identical across domains due to the discriminative information from labels. We are motivated to propose a novel deep network $h(\mathbf{x})$ that formally reduces the data distribution shift across domains and achieves better performance on the target domain, such that the target risk $\epsilon_t(h) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_t[h(\mathbf{x})\neq\mathbf{y}]}$ can be safely bounded by using the source domain.

B. Network Architecture

Inspired by GAN, adversarial domain adaptation networks [40] have been investigated to integrate domain adaptation and adversarial learning to learn domain-invariant features in a two-player game. We are further motivated to propose a novel adversarial domain adaptation network to simultaneously align the marginal and conditional distributions for feature representation. Then, the classifier trained on the annotated source data can be safely applied to predict labels for the target data.

Fig. 2 shows the overall architecture of our proposed DJDAN model. Firstly, EEG signals $\mathbf{x}^s$ and $\mathbf{x}^t$ are transformed into high-level discriminative features $\mathbf{f}^s$ and $\mathbf{f}^t$ by

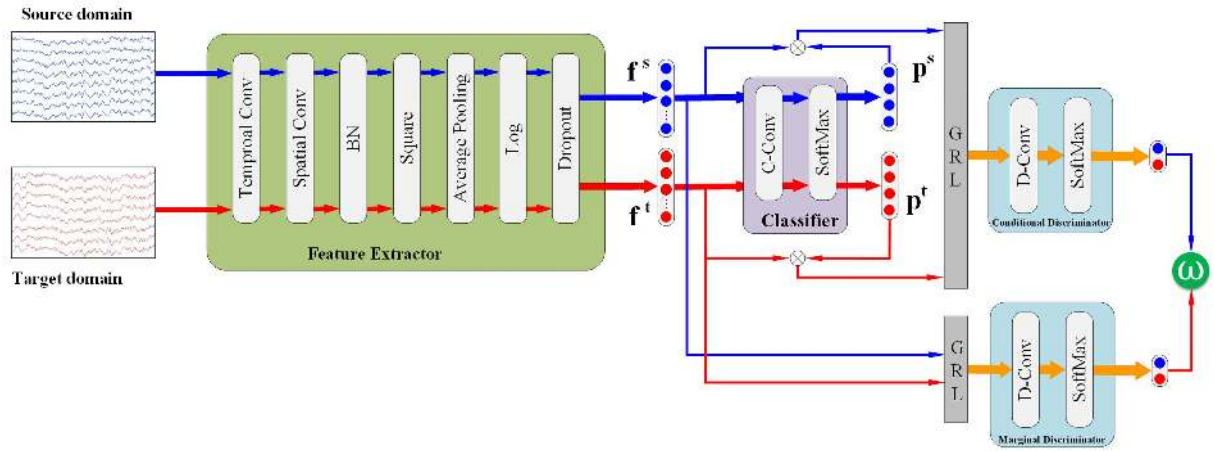

Fig. 2. The architecture of the proposed DJDAN model, which consists of four components, including the feature extractor, the classifier, the marginal discriminator and the conditional discriminator. ⊗ denotes the product operator, f denotes the feature representation and p represents the predicted labels.

the feature extractor F for both the source and target domains, respectively. Then, these features are subsequently fed into the classifier G to obtain the corresponding predicted labels. Based on the extracted features and predicted labels, the adaptation of marginal and conditional distribution is achieved by the global domain discriminator D_G and the local domain discriminator D_L. The global domain discriminator pays attention to the marginal distributions and is trained to recognize which domain the high-level features come from. Meanwhile, the local domain discriminator considers the local multi-modal structures and recognizes the data domains with assistance of discriminative information conveyed in the predicted labels. To measure the importance between the global and local distribution alignment, we also utilize a dynamic adversarial factor ω to evaluate the weight between the global and local discriminators during training. In the adversarial procedure, D_G and D_L, as the first player, are trained to distinguish the source domain from the target domain *w.r.t*, the global and local distributions. Then, as the second player, the feature extractor and the classifier are jointly trained to learn global and local domain-invariant features, and thus confuse the first player. In the following, we will introduce the feature extractor, classifier, marginal discriminator and conditional discriminator in details.

1) Feature Extractor: Different from images, the multi-electrode EEG signal $\mathbf{x} \in \mathbb{R}^{E\times T}$ as the network input has inconsistent dimension units in the spatial and temporal dimensions. Thus, it results in a non-trivial choice of convolutional network and kernel sizes. Similar to the shallow ConvNet [19], our feature extractor also employs two one-dimensional convolutional operations, namely, temporal and spatial convolutions for feature extraction. The first layer performs temporal convolution along the time axis. It learns temporal and frequency information with 40 kernels of length 25, which is analog to the electrode-wise band-pass filtering operations in previous works [15]. Then, a spatial convolution layer is connected to learn the spatial representation along the electrode axis, where the kernel length is equal to the number of electrodes. The extracted features of spatial convolution are fed into a batch normalization (BN) layer before the squaring activation to avoid gradient vanishing problem. The output is subsequently connected to an average pooling layer, logarithm activation and dropout layer to prevent over-fitting. The details of hyper-parameters are listed in Table I. As a result, we obtain a group of high-level feature representation $\mathbf{f}^s$ and $\mathbf{f}^t$ for source and target domain, respectively.

2) Classifier: The classifier follows the feature extractor module, and is connected to high-level features $\mathbf{f}^s$ and $\mathbf{f}^t$ to learn the label prediction. Specifically, $\mathbf{f}^s$ and $\mathbf{f}^t$ are fed into a fully convolutional layer with the number of output neurons being equivalent to the task categories C. Then, a softmax operation is employed to transform the output results into probability estimate for each category. Since only the source EEG data are annotated, the classifier and feature extractor are trained on the source data with:

$$L_c(\theta_f, \theta_c) = \mathbb{E}_{(\mathbf{x}_i^s, \mathbf{y}_i^s)\sim \mathcal{D}_s} L(\mathbf{p}_i^s, \mathbf{y}_i^s), \tag{1}$$

where θ_f and θ_c denote the model parameters in the feature extractor and classifier, respectively, $\mathbf{p}_i^s$ is the conditional probability vector generated by the softmax function, $\mathbf{y}_i^s$ is the corresponding label, and $L(\cdot)$ is the cross-entropy loss function.

3) Global Discriminator: In BCI applications, features generated from the feature extractor may have different marginal distributions across domains. To distinguish which domain the features are generated from, D_G performs a binary classification in a supervised way. Specifically, the global discriminator sets the domain labels of target features to be 1 and those of source features to 0. Similar to [40], we calculate the loss of the global discriminator with

$$\begin{aligned} L_g(\theta_f, \theta_g) = &-\mathbb{E}_{\mathbf{x}^s\sim \mathcal{D}_s} \log[D_G(F(\mathbf{x}^s))] \\ &-\mathbb{E}_{\mathbf{x}^t\sim \mathcal{D}_t} \log[1 - D_G(F(\mathbf{x}^t))], \end{aligned} \tag{2}$$

where $F(\mathbf{x}^s)$ and $F(\mathbf{x}^t)$ denote the feature extractor in the source and the target domain, respectively.

4) Local Discriminator: With the global discriminator, the marginal distributions between the source and target

TABLE I
MODEL HYPER-PARAMETERS OF DJDAN

Modules	Layers	Parameters	Output
Input	-	-	1×E×T
Feature extractor F	Temporal Conv	1×25, 40	$40 \times E \times 976$
	Spatial Conv	$E \times 1$, 40	$40 \times 1 \times 976$
	BN	-	$40 \times 1 \times 976$
	Square activaction	-	$40 \times 1 \times 976$
	Average Pooling	1×75, stride 15	$40 \times 1 \times 61$
	Logarithm activaction	-	$40 \times 1 \times 61$
	Dropout	$p = 0.5$	$40 \times 1 \times 61$
Classifier G	C-Conv	1×61, C	$C \times 1$
	Softmax	-	$C \times 1$
Marginal discriminator D_G	D-Conv	1×61, 2	2×1
	Softmax	-	2×1
Conditional discriminator D_L	D-Conv	1×61, 2	2×1
	Softmax	-	2×1

domains are drawn closer. However, reducing the marginal distribution discrepancy does not guarantee that the conditional distributions are aligned, as shown in Fig. 1 (*Source Domain* vs. *Target Domain II*). Thus, we design the local discriminator module to align conditional distributions such that $P(\mathbf{x}^s|\mathbf{y}^s) \approx P(\mathbf{x}^t|\mathbf{y}^t)$. This problem is nontrivial, since there are no labels $\mathbf{y}^t$ available in the target domain. Inspired by [41], we assume the labels predicted by the classifier G may contain potential discriminative information, which can be leveraged to perform a fine-grained domain adaptation.

Here, the local discriminator D_L is conditioned on the classifier prediction, and is split into C class-wise domain discriminators D_L^c, each is associated with class $c \in \{1, 2, \cdots, C\}$. Specifically, the classifier prediction can be used to indicate the confidence that the feature representation should belong to each of C categories, which is estimated by the corresponding conditional domain discriminator. Similar to the global discriminator, the loss of the local discriminator for the c^{th} category is calculated with

$$L_l^c(\theta_f, \theta_l) = -\mathbb{E}_{\mathbf{x}^s \sim \mathcal{D}_s} log[D_L(\mathbf{p}_c^s F(\mathbf{x}^s))] - \mathbb{E}_{\mathbf{x}^t \sim \mathcal{D}_t} log[1 - D_L(\mathbf{p}_c^t F(\mathbf{x}^t))], \quad (3)$$

where $\mathbf{p}_c^s$ and $\mathbf{p}_c^t$ denote the predicted probability distribution of the input sample $\mathbf{x}^s$ and $\mathbf{x}^t$ belonging to the c^{th} class in the source and the target domain, respectively. The simple combination of predicted probability and feature representation $\mathbf{p}_c F(\mathbf{x})$ explores discriminative information embedded in the multi-modal structures [41]. In this regard, each conditional domain discriminator focuses on the local distributions between the source and the target domains. The loss of the local discriminator can be calculated as the sum of $L_l^c(\theta_f, \theta_l)$ for all C categories.

C. Dynamic Adversarial Factor

The global and the local discriminators may make different contributions to domain adaption, which is adjusted by the weight ω between their loss functions. Instead of random guessing using a fixed weight in range [0, 1], we introduce a *dynamic adversarial factor* ω to easily, dynamically, and quantitatively evaluate the relative importance of the marginal and conditional distribution alignment. Concretely, we employ the $\mathcal{A}$-distance [42] to measure the marginal distribution and conditional distribution discrepancies across domains. Here, we denote the global $\mathcal{A}$-distance for the global discriminator with

$$d_{\mathcal{A},g}(\mathcal{D}_s, \mathcal{D}_t) = 2(1 - 2\epsilon_g), \quad (4)$$

where ϵ_g represents the classification error rate for the global domain discriminator. Similarly, the local $\mathcal{A}$-distance of the local discriminator over the c^{th} class is represented with

$$d_{\mathcal{A},l}(\mathcal{D}_s^c, \mathcal{D}_t^c) = 2(1 - 2\epsilon_l^c), \quad (5)$$

where $\mathcal{D}_s^c$ and $\mathcal{D}_t^c$ denote samples from the c^{th} class and ϵ_l^c is the classification error of local sub-domain discriminator loss over class c.

Then, with the global and local $\mathcal{A}$-distances, the dynamic adversarial factor ω is calculated as

$$\omega = \frac{d_{\mathcal{A},g}(\mathcal{D}_s, \mathcal{D}_t)}{d_{\mathcal{A},g}(\mathcal{D}_s, \mathcal{D}_t) + \frac{1}{C}\Sigma_{c=1}^{C} d_{\mathcal{A},l}(\mathcal{D}_s^c, \mathcal{D}_t^c)}. \quad (6)$$

For initialization, ω is set to be 0.5 in the first epoch. In the subsequent epochs, the dynamic adversarial factor can be estimated with the pseudo labels predicted by the classifier. During training, if the $\mathcal{A}$-distance is larger in global discriminator, ω would be larger and thus drives our DJDAN to pay more attention to the global distribution alignment, vice versa. Eventually, our DJDAN will learn a rather robust dynamic adversarial factor.

D. Optimization of Network

During training, the proposed DJDAN is jointly optimized with three components, namely, the classifier loss (Eq. (1)),

the global discriminator loss (Eq. (2)), and the local discriminator loss (Eq. (3)). The overall loss function can be finally formulated with

$$L(\theta_f, \theta_c, \theta_g, \theta_l) = L_c - \alpha(\omega L_g + (1-\omega)\sum_{c=1}^{C} L_l^c), \quad (7)$$

where α is the trade-off hyper-parameter between the classifier and the discriminators, and ω is the dynamic adversarial factor, which is adaptively updated during training.

Generally, to confuse the domain discriminators and learn domain-invariant features, the optimal parameters $\hat{\theta}_f$, $\hat{\theta}_c$, $\hat{\theta}_g$ and $\hat{\theta}_l$ can be learned by alternately minimizing and maximizing the loss function of $L(\theta_f, \theta_c, \theta_g, \theta_l)$ in Eq. (7). Firstly, we update the parameters of θ_f and θ_c by minimizing the loss function as follows:

$$(\hat{\theta}_f, \hat{\theta}_c) = arg\min_{\theta_f, \theta_c} L(\theta_f, \theta_c, \hat{\theta}_g, \hat{\theta}_l). \quad (8)$$

Then, after obtaining the optimal values of $\hat{\theta}_f$ and $\hat{\theta}_c$, the optimal values of $\hat{\theta}_g$ and $\hat{\theta}_l$ can be updated by maximizing the following function:

$$(\hat{\theta}_g, \hat{\theta}_l) = arg\max_{\theta_g, \theta_l} L(\hat{\theta}_f, \hat{\theta}_c, \theta_g, \theta_l). \quad (9)$$

As a result, the feature extractor will generate the feature representations, which can minimize the loss of classifier and maximize the loss of domain discriminator simultaneously. When the trained optimal discriminator cannot distinguish whether features come from the source domain or target domain, we obtain the common motor-imagery EEG features that exist in both the source and target domains.

For implementation, we adopt a gradient reversal layer (GRL) [40] that acts like an identity layer in the forward propagation and reverses the gradients in the back-propagation stage by multiplying the gradient with -1. In this way, the parameters learned in the feature extractor essentially perform gradient ascent with respect to the gradients in the domain discriminators.

E. Data Preprocessing

Our proposed DJDAN is capable of learning discriminative representation and only requires two simple preprocessing operations before feeding the EEG data into the network, namely, band-pass filtering and exponential moving standardization. For frequency filtering, a third-order Butterworth band-pass filter of [4−38] Hz is conducted to remove unrelated information with respect to the motor imagery task from the raw EEG trials. With filtered EEG signals, we further employ electrode-wise exponential moving standardization to eliminate undesirable noises with

$$x_k = \frac{\widetilde{x}_k - \mu_k}{\sqrt{\sigma_k^2}}, \quad (10)$$

where x_k and $\widetilde{x}_k$ denote the standardized and input filtered signal at time k, respectively. μ_k and σ_k are the corresponding exponential mean value and variance formulated as

$$\mu_k = (1-\beta)\widetilde{x}_k + \beta\mu_{k-1}, \quad (11)$$

$$\sigma_k^2 = (1-\beta)(\widetilde{x}_k - \mu_k)^2 + \beta\sigma_{k-1}^2, \quad (12)$$

where β is a decay factor and is set to be 0.999. Note that with these two simple preprocessing operations, the resulting EEG signals not only preserve useful motor imagery information, but also eliminate occasional noises, which can be safely fed into our network.

F. Generalization Error

According to the domain adaption theory [41], we provide an analysis of the expected risk of the proposed DJDAN model. Considering a family of source classifiers h in the hypothesis space $\mathcal{H}$, we denote the expected risk on the source domain as $\epsilon_s(h, f_s) = \mathbb{E}_{\mathbf{x}_s \sim \mathcal{D}_s}[h(\mathbf{x}_s) \neq f_s(\mathbf{x}_s)]$, where f_s is the labeling function of the source domain. Similarly, $\epsilon_t(h, f_t) = \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[h(\mathbf{x}_t) \neq f_t(\mathbf{x}_t)]$ is the expected risk on the target domain w.r.t. distribution $\mathcal{D}_t$, which can be bounded by the following inequality [42]

$$\epsilon_t(h, f_t) \leq \epsilon_s(h, f_s) + |\epsilon_t(h, f_s) - \epsilon_s(h, f_s)| + |\epsilon_t(h, f_t) - \epsilon_t(h, f_s)|. \quad (13)$$

The goal of domain adaptation is to reduce the marginal distribution discrepancy $|\epsilon_t(h, f_s) - \epsilon_s(h, f_s)|$ and conditional distribution discrepancy $|\epsilon_t(h, f_t) - \epsilon_t(h, f_s)|$. By definition, we have

$$|\epsilon_t(h, f_t) - \epsilon_t(h, f_s)| = \Big|\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[h(\mathbf{x}_t) - f_t(\mathbf{x}_t) \neq 0] - \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[h(\mathbf{x}_t) - f_s(\mathbf{x}_t) \neq 0]\Big|, \quad (14)$$

and

$$|\epsilon_t(h, f_s) - \epsilon_s(h, f_s)| = \Big|\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[h(\mathbf{x}_t) - f_s(\mathbf{x}_t) \neq 0] - \mathbb{E}_{\mathbf{x}_s \sim \mathcal{D}_s}[h(\mathbf{x}_s) - f_s(\mathbf{x}_s) \neq 0]\Big|. \quad (15)$$

Then, we define a loss hypothesis space $\triangle = \{\delta(f) = |h(\mathbf{x}) - f(\mathbf{x})|, f \in \mathcal{H}\}$ over the labeling function f, where $\delta \mapsto \{0, 1\}$ outputs the distance between any h and a specific labeling function f. According to the above loss hypothesis space $\triangle$, we define the $\triangle$-distance as following

$$d_\triangle = \Big|\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[\delta(f_t) \neq 0] - \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[\delta(f_s) \neq 0]\Big|. \quad (16)$$

Consequently, the conditional distribution discrepancy $|\epsilon_t(h, f_t) - \epsilon_t(h, f_s)|$ can be upper-bounded by the $\triangle$-distance. Since multilayer perceptron can fit any functions, the family of domain discriminators $\mathcal{H}_D$ is rich enough to contain the loss hypothesis space $\triangle$. Then, the proposed local discriminator D_L is related to $d_\triangle$:

$$\begin{aligned} d_\triangle &\leq \sup_{D_L \in \mathcal{H}_D} \Big|\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[D_L(f_t) \neq 0] - \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[D_L(f_s) \neq 0)]\Big| \\ &\leq \sup_{D_L \in \mathcal{H}_D} \Big|\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[D_L(f_t) = 1] - \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t}[D_L(f_s) = 0]\Big|. \end{aligned} \quad (17)$$

Similarly, the training marginal discriminator D_G is also related to $d'_{\triangle}$:

$$\begin{aligned} d'_{\triangle} &= \left|\mathbb{E}_{\mathbf{x}_t\sim\mathcal{D}_t}[\delta(f_s)\neq 0] - \mathbb{E}_{\mathbf{x}_s\sim\mathcal{D}_s}[\delta(f_s)\neq 0]\right| \\ &\leq \sup_{D_G\in\mathcal{H}_D} \left|\mathbb{E}_{\mathbf{x}_t\sim\mathcal{D}_t}[D_G(\mathbf{x}_t)\neq 0] - \mathbb{E}_{\mathbf{x}_s\sim\mathcal{D}_s}[D_G(\mathbf{x}_s)\neq 0]\right| \\ &\leq \sup_{D_G\in\mathcal{H}_D} \left|\mathbb{E}_{\mathbf{x}_t\sim\mathcal{D}_t}[D_G(\mathbf{x}_t)= 1] - \mathbb{E}_{\mathbf{x}_s\sim\mathcal{D}_s}[D_G(\mathbf{x}_s)= 0]\right|. \end{aligned} \tag{18}$$

The above suprema are achieved in the process of training the optimal marginal discriminator D_G and conditional discriminator D_L in DJDAN, giving an upper bound of $d_{\triangle}$ and $d'_{\triangle}$, respectively.

IV. Experiments

A. Datasets

To evaluate our proposed DJDAN model, we conduct extensive experiments on two public motor imagery datasets, namely, Dataset IIa and IIb of BCI Competition IV.

1) Dataset IIa of BCI Competition IV: This dataset [46] collected 22-electrode EEG signals in two different sessions from nine healthy subjects (refer to $A1-A9$). Each subject participated in four motor imagery tasks, including imagining the movement of left hand, right hand, feet and tongue. It contains 72 trials of EEG data for each task in both sessions. In this paper, we regard EEG data in the first session as the training data (source domain) and those in the second session as the test one (target domain). Note that the time segment between $[2, 6]$ second for each trial is considered in our experiments.

2) Dataset IIb of BCI Competition IV: This dataset [47] recorded 3-electrode EEG motor-imagery signals in five sessions from nine subjects (refer to $B1-B9$). Each participant performed binary-category movement imagery tasks, namely, left hand or right hand. Similar to [43], we also select the first three sessions as the training data (source domain) and the rest for test (target domain). Then, we have about 400 trials of EEG data in the source domain and about 320 trials in the target domain. Note that we consider the time segment of $[3, 7]$ second for our experiments.

B. Experiment Settings

To demonstrate the advantages of our method, we compare the performance of our method with the following state-of-the-art algorithms, including FBCSP [15], CCSP [22], SSMM [43], ConvNet [19], C2CM [44], and EEGNet [45]. Specifically, FBCSP [15], based on the filter-bank common spatial patterns, was the winner algorithm for both datasets in BCI Competiton IV. SSMM [43] is an efficient matrix classifier involving two-dimensional data, like EEG features. ConvNet [19] is a shallow convolutional neural network tailored to decode band power features. C2CM [44] is a deep convolutional neural network, which fine-tunes the network hyper-parameters for each subject, such as hidden nodes and kernel size. EEGNet [45] is a compact CNN framework designed for EEG signals decoding, such as P300 event-related potential (P300), movement-related cortical potential (MRCP), motor imgery and so on. CCSP [22] is a modified CSP-based method for subject-to-subject transformation, which determines the composite covariance matrices by a weighted sum of covariance matrices from all subjects. For fair comparison, we follow the evaluation protocol in unsupervised domain adaptation [48], and select the annotated training data as source data and test data without labels as target data for each subject in both datasets.

We implement our approach in PyTorch with an Intel Core I7 CPU and a Tesla P40 GPU. For both datasets, EEG signals from all electrodes are used for classification and the three electrooculography (EOG) channels are directly discarded without any artifact removing operation. The proposed DJDAN model is trained from scratch via back-propagation with batch size of 64. We adopt the Adam optimizer with momentum of 0.9 and the learning rate of 0.0005. α is set to be 0.3 during training. We also employ an early stop strategy [49] to terminate the model training if no loss descent is observed in 20 steps to avoid over-fitting.

C. Experimental Results

We evaluate different algorithms on the Dataset IIa and report the classification accuracy for each subject and the average accuracy in Table II. As is shown, the proposed method outperforms all state-of-the-art algorithms with a large margin. It demonstrates that our dynamic joint domain adaptation network is capable of reducing both the marginal and conditional discrepancies across domains, and thus efficiently improves the performance in the target domain. From the experimental results, we have the following observations. Firstly, neural network based methods (ConvNet, C2CM, EEGNet and ours) can achieve comparable performance and even outperform those traditional methods, such as FBCSP and CCSP. It illustrates that deep neural networks are capable of learning the discriminative features for EEG classification. Secondly, C2CM shows superior performance compared with ConvNet, implying that fine-tuning the architecture parameters for each subject may improve the classification performance. However, such fine-tuning strategy is time consuming in real-world applications. Thirdly, different from our domain adaptation protocol, CCSP only reduce the marginal domain discrepancy by using data from other subjects, and show inferior classification performance to ours. It indicates that domain discrepancy is difficult to be reduced by using different subjects' data, and inaccurate weight estimation of previous subjects may result in the negative adaptation as their data are not aligned properly. Moreover, it may not guarantee similar conditional distribution even with close marginal distribution, which could deteriorate the classification performance of models.

For further verifying the effectiveness of our method, the results on the Dataset IIb are reported in Table III. It is noteworthy that our proposed method greatly improves the classification accuracy compared with the state-of-the-art methods. It demonstrates that our framework is effective for EEG decoding and classification. Compared with ConvNet, our method further achieves an average 3.63% improvement, implying that the joint global and local domain adaptation is

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

Method	Subject									Average ACC
	A1	A2	A3	A4	A5	A6	A7	A8	A9	
FBCSP [15]	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75
CCSP [22]	84.72	52.78	80.90	59.38	54.51	49.31	88.54	71.88	56.60	66.51
SSMM [43]	82.64	60.76	85.76	67.01	58.68	54.51	90.97	81.25	79.51	73.45
ConvNet [19]	76.39	55.21	89.24	74.65	56.94	54.17	92.71	77.08	76.39	72.53
C2CM [44]	87.50	65.28	90.28	66.67	62.50	45.49	89.58	83.33	79.51	74.46
EEGNet [45]	85.76	61.46	88.54	67.01	55.90	52.08	89.58	83.33	**86.81**	74.50
Ours	**86.46**	**68.75**	**93.06**	**85.42**	**72.57**	**63.54**	**95.49**	**85.76**	83.68	**81.52**

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

Method	Subject									Average ACC
	B1	B2	B3	B4	B5	B6	B7	B8	B9	
FBCSP [15]	70.00	**60.36**	60.94	97.50	93.12	80.63	78.13	92.50	86.88	80.00
CCSP [22]	63.75	56.79	50.00	93.44	65.63	81.25	72.81	87.81	82.81	72.70
SSMM [43]	74.06	55.00	55.63	94.06	86.88	82.19	76.56	92.19	85.62	78.00
ConvNet [19]	76.56	50.00	51.56	96.88	93.13	**85.31**	83.75	91.56	85.62	79.37
EEGNet [45]	68.44	57.86	**61.25**	90.63	80.94	63.13	84.38	**93.13**	83.13	75.88
Ours	**83.44**	58.57	59.06	**98.13**	**96.56**	84.38	**86.25**	92.81	**87.81**	**83.00**

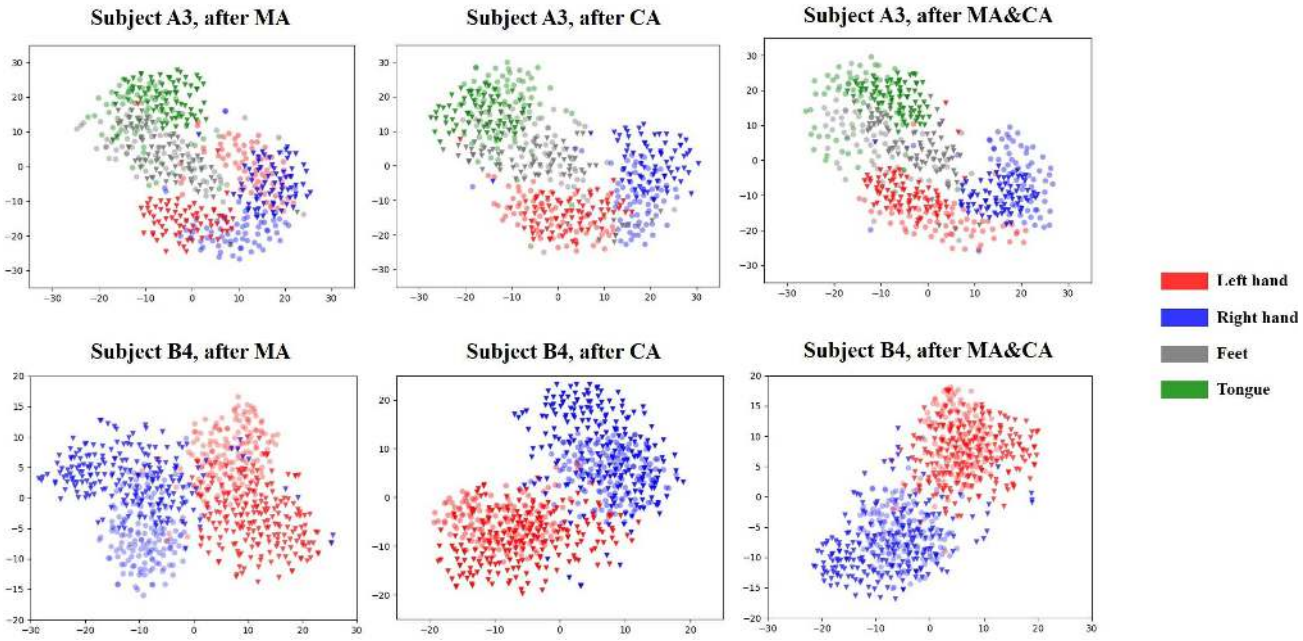

Fig. 3. Feature visualization by t-SNE. Triangles denote deep features from the source domain and dots represent features from the target domain. The first column shows the features after marginal alignment (MA), the second column shows the features after conditional alignment (CA), and the third after MA and CA.

helpful to reduce the distribution discrepancy across domains and leverage the useful information from the source domain.

D. Effectiveness Analysis

Ablation Study We compare the performance of two variants of the proposed DJDAN: (1) DJDAN using only marginal discriminator for marginal distribution alignment (MA), referred to MAAN ($\omega = 0$). (2) DJDAN with only conditional discriminator for conditional distribution alignment (CA), termed CAAN ($\omega = 1$). Both MAAN and CAAN can be regarded as the special cases of our DJDAN. The average results on both datasets are reported in Fig. 4. From the results,

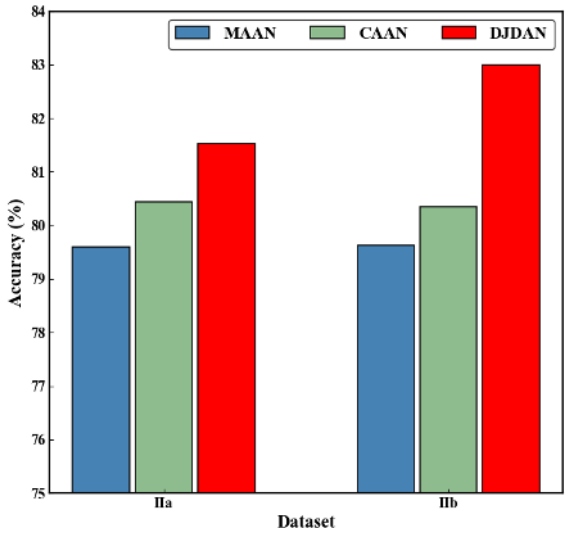

Fig. 4. Ablation study of DJDAN on datasets IIa and IIb.

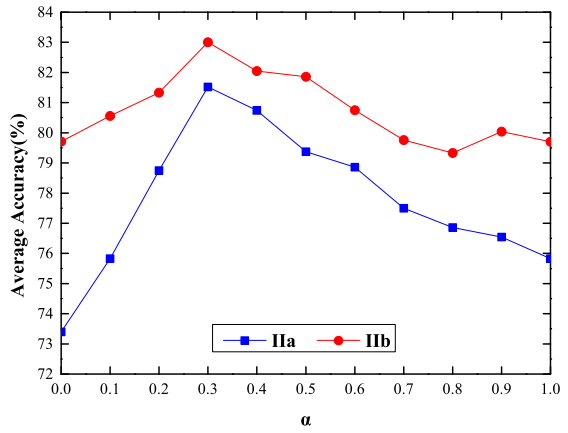

Fig. 5. The sensitivity of average accuracy of DJDAN to α on datasets IIa and IIb.

DJDAN achieves the highest classification accuracy on both datasets compared with MAAN and CAAN. It indicates that it is not enough to only align the marginal or conditional distributions, and simultaneously considering both global and local distribution alignment can achieve better performance.

Feature Visualizations Then, we are also interested in exploring the distributions of feature representation learned by MAAN, CAAN and our DJDAN. We randomly select subjects A3 and B4 from these two datasets and visualize their feature representation learned by MAAN, CAAN and DJDAN using t-SNE embeddings [50], which are presented in Fig. 3. The triangles represent features from the previous session (source domain), and the dots from the current session for same subject (target domain). For a better visualization, we also highlight features from different categories with different colors. The visualization shows several interesting observations. (1) Though the global distribution discrepancy is reduced with marginal adaptation, the local distributions may still be very different across domains, like the "left hand vs. right hand" features for subject A3 in Dataset IIa. (2) The features learned with CAAN for subject A3 show that local distributions across domains may be well aligned, implying the conditional domain discrepancy is significantly reduced after conditional alignment. However, the marginal distributions of "right hand" for subject B4 are still different between the source and target domains. (3) Compared with MAAN and CAAN features, our method achieves better feature alignment of both global and local distributions, via simultaneously reducing the marginal and conditional distribution discrepancies. In this regard, our method is capable of learning the discriminative and domain-invariant features, leading to robust and superior classification performance.

Parameter Sensitivity We further investigate the impact of the hyper-parameter α in Eq. (7). Fig. 5 gives an illustration of the variation of transfer classification performance as $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ on Dataset IIa and IIb. We can observe that the classification accuracy demonstrates a bell-shaped curve as α varies from 0 to 1. This also verifies the efficacy of jointly learning deep features and adapting distribution discrepancy, since a good trade-off between them can enhance feature transferability. Specially, DJDAN is less sensitive to the change of α in Dataset IIb.

V. Conclusion

In this paper, we have proposed a novel dynamic joint domain adaptation neural network, referred to DJDAN, to extract more transferable features for cross-session motor imagery classification. Different from traditional EEG classification methods, our DJDAN model has explored a deep architecture to learn the discriminative features in an end-to-end manner. In addition, to learn domain-invariant features from the multi-modal structures, our method simultaneously reduced the marginal and conditional distribution discrepancies across domains via the global and local discriminators. Moreover, we have investigated an adversarial factor ω to dynamically evaluate the importance between the global and local distribution adaptation. Finally, the extensive experimental results have demonstrated that our method could achieve superior classification performance compared with state-of-the-art methods.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] L. Bi, J. Zhang, and J. Lian, "EEG-based adaptive driver-vehicle interface using variational autoencoder and PI-TSVM," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2025–2033, Oct. 2019.

[3] J.-H. Jeong, K.-H. Shim, D.-J. Kim, and S.-W. Lee, "Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1226–1238, May 2020.

[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.*, early access, Dec. 6, 2019, doi: 10.1109/TSMC.2019.2956022.

[5] L. Yao *et al.*, "A stimulus-independent hybrid BCI based on motor imagery and somatosensory attentional orientation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1674–1682, Sep. 2017.

[6] A. E. Hassanien and A. Azar, *Brain-Computer Interfaces*, vol. 74. Cham, Switzerland: Springer, 2015.

[7] W. Zheng, "Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis," *IEEE Trans. Cognit. Develop. Syst.*, vol. 9, no. 3, pp. 281–290, Sep. 2017.

[8] Z. Gao *et al.*, "EEG-based spatio–temporal convolutional neural network for driver fatigue evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2755–2763, Sep. 2019.

[9] 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.

[10] L. S. Vidyaratne and K. M. Iftekharuddin, "Real-time epileptic seizure detection using EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 2146–2156, Nov. 2017.

[11] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: Applying brain–computer interface technology to human–machine systems in general," *J. Neural Eng.*, vol. 8, no. 2, Apr. 2011, Art. no. 025005.

[12] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier," *Comput. Methods Programs Biomed.*, vol. 108, no. 1, pp. 10–19, Oct. 2012.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Cham, Switzerland: Springer, 2006.

[14] Z. Gao, X. Cui, W. Wan, and Z. Gu, "Recognition of emotional states using multiscale information analysis of high frequency EEG oscillations," *Entropy*, vol. 21, no. 6, p. 609, Jun. 2019.

[15] 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.

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

[17] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[18] 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.

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

[20] A. M. Azab, J. Toth, L. S. Mihaylova, and M. Arvaneh, "A review on transfer learning approaches in brain–computer interface," in *Signal Processing and Machine Learning for Brain-Machine Interfaces*. London, U.K.: Institution of Engineering and Technology, 2018, pp. 81–98.

[21] 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.

[22] 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.

[23] 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.

[24] 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)*, Feb. 2019, pp. 1–4.

[25] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Dec. 7, 2019, doi: 10.1109/TAFFC.2018.2885474.

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

[27] H.-J. Kim, I.-N. Wang, Y.-T. Kim, H. Kim, and D.-J. Kim, "Comparative analysis of NIRS-EEG motor imagery data using features from spatial, spectral and temporal domain," in *Proc. 8th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2020, pp. 1–4.

[28] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Nov. 2013, pp. 81–84.

[29] Y. Yang, Q. M. J. Wu, W.-L. Zheng, and B.-L. Lu, "EEG-based emotion recognition using hierarchical network with subnetwork nodes," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 2, pp. 408–419, Jun. 2018.

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

[31] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki, "Correlation-based channel selection and regularized feature optimization for MI-based BCI," *Neural Netw.*, vol. 118, pp. 262–270, Oct. 2019.

[32] J. Jin, R. Xiao, I. Daly, Y. Miao, X. Wang, and A. Cichocki, "Internal feature selection method of CSP based on L1-norm and dempster-shafer theory," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 24, 2020, doi: 10.1109/TNNLS.2020.3015505.

[33] H. Dose, J. S. Møller, 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] 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. Ind. Informat.*, vol. 16, no. 11, pp. 7159–7167, Nov. 2020.

[35] 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.*, early access, Feb. 4, 2020, doi: 10.1109/TSMC.2020.2964684.

[36] Z. Gao *et al.*, "Complex networks and deep learning for EEG signal analysis," *Cogn. Neurodyn.*, 2020, doi: 10.1007/s11571-020-09626-1.

[37] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *Ann. Statist.*, vol. 41, no. 5, pp. 2263–2291, Oct. 2013.

[38] X. Tang and X. Zhang, "Conditional adversarial domain adaptation neural network for motor imagery EEG decoding," *Entropy*, vol. 22, no. 1, p. 96, Jan. 2020.

[39] C. Yu, J. Wang, Y. Chen, and M. Huang, "Transfer learning with dynamic adversarial adaptation network," 2019, *arXiv:1909.08184*. [Online]. Available: http://arxiv.org/abs/1909.08184

[40] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.

[41] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1640–1650.

[42] 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, nos. 1–2, pp. 151–175, May 2010.

[43] 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.

[44] 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.

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

[46] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set A," Inst. Knowl. Discovery, Lab. Brain-Comput. Interfaces, Graz Univ. Technol., Graz, Austria, Tech. Rep., 2008, pp. 136–142.

[47] 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., 2008, pp. 1–6.

[48] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

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

[50] J. Donahue *et al.*, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.