Domain Adaptation for EEG Emotion Recognition Based on Latent Representation Similarity

Jinpeng Li, Shuang Qiu, Changde Du, Yixin Wang, and Huiguang He, *Senior Member, IEEE*

***Abstract*—Emotion recognition has many potential applications in the real world. Among the many emotion recognition methods, electroencephalogram (EEG) shows advantage in reliability and accuracy. However, the individual differences of EEG limit the generalization of emotion classifiers across subjects. Moreover, due to the nonstationary characteristic of EEG, the signals of one subject change over time, which is a challenge to acquire models that could work across sessions. In this article, we propose a novel domain adaptation method to generalize the emotion recognition models across subjects and sessions. We use neural networks to implement the emotion recognition models, which are optimized by minimizing the classification error on the source while making the source and the target similar in their latent representations. Considering the functional differences of the network layers, we use adversarial training to adapt the marginal distributions in the early layers and perform association reinforcement to adapt the conditional distributions in the last layers. In this way, we approximately adapt the joint distributions by simultaneously adapting marginal distributions and conditional distributions. The method is compared with multiple representatives and recent domain adaptation algorithms on benchmark SEED and DEAP for recognizing three and four affective states, respectively. The experimental results show that the proposed method reaches and outperforms the state of the arts.**

***Index Terms*—Domain adaptation, electroencephalogram (EEG), emotion recognition, neural network, transfer learning.**

I. Introduction

EMOTION recognition is an important issue in both scientific research and engineering [1]. In the psychology research, it provides a quantitative reference for studying emotion-related behaviors. In engineering, it facilitates a more friendly human–machine interaction, where machines recognize, understand, and even generate human emotions [2]. In the medical field, it helps the diagnosis and treatment of various mental diseases, such as autism spectrum disorders [3] and depression [4]. In addition, the assessment of emotion also helps doctors to track the recovery of patients. The basis of emotion recognition is the physiological or nonphysiological signals distributed across the human body. As a kind of physiological signal, electroencephalogram (EEG) is more objective and reliable than nonphysiological approaches, such as facial expression, gesture, and language [5]. Therefore, in recent years, EEG emotion recognition has attracted strong interests among researchers and engineers [6].

In EEG emotion recognition, the paradigm design, signal preprocessing, feature extraction, and classification methods have been intensively investigated [6]–[8]. However, there are still some challenges. Due to the individual differences and the nonstationary characteristic of EEG [9], the models are hard to generalize across subjects and easily become obsolete over time. We have to collect labeled samples for each subject at each time to train new models. However, labeling work is time consuming and expensive. To reduce the reliance on the labeled data, this article generalizes models across subjects and sessions with the domain adaptation approach. We refer to the EEG of different subjects or sessions as individual domains. In the cross-subject case, the target (domain) refers to the new subject, and the source (domain) refers to an existing subject. In the cross-session case, the target is the new subject, and the source is his/her existing sessions. In both cases, there might be multiple sources. The source is labeled while the target is unlabeled. We use source labels to deduce target labels. Seldom works have discussed this topic in EEG study [10].

We adapt marginal and conditional distributions of the source and target simultaneously to encourage similarity in joint distributions. Considering the functional differences of network layers, we adapt marginal distributions at shallower layers who produce task-invariant features, and conditional distributions at deeper layers who produce task-specific features. As the variability of the two domains is reduced, source supervisions can be used in the target. The contribution of this article is twofold: 1) we propose an effective joint distribution adaptation (JDA) method and 2) we exploit the connection

Manuscript received September 29, 2018; revised March 5, 2019 and July 30, 2019; accepted October 10, 2019. Date of publication October 24, 2019; date of current version June 10, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61976209 and Grant 81701785, in part by Chinese Academy of Sciences (CAS) International Collaboration Key Project, and in part by the Strategic Priority Research Program of CAS under Grant XDB32040200. *(Corresponding author: Huiguang He.)*

J. Li is with the Research Center for Brain-Inspired Intelligence and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China, and also with the Ningbo Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo 315010, China (e-mail: lijinpeng.ai@hotmail.com).

S. Qiu, C. Du, and Y. Wang are with the Research Center for Brain-Inspired Intelligence and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China

H. He is with the Research Center for Brain-Inspired Intelligence and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: huiguang.he@ia.ac.cn).

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/TCDS.2019.2949306

between adaptation strategies and functional layers of neural networks. The experimental results show the superiority of our method in both cross-subject and cross-session cases.

The rest of this article is organized as follows. Section II is a review of EEG emotion recognition and domain adaptation algorithms. Section III presents the general framework and the mathematical descriptions of our method. In Section IV, we evaluate the method on two open data sets. Section V discusses and analyzes the results. Section VI is the conclusion.

II. Related Work

From the perspective classification model in EEG study, until now, most works used support vector machines (SVMs). Other popular models include k-nearest neighbor (kNN) and linear discriminant analysis (LDA). These shallow models use EEG features directly to conduct classification [6]. In recent years, researchers introduced deep neural networks (DNNs) to emotion recognition, which showed an advantage over shallow models due to the representational learning capacity [6], [8], [11]. Although DNN has shown some successful applications, the major challenging to EEG remains unsolved, i.e., EEG signals are temporal asymmetry and nonstationary. This means the data distribution is different across subjects and sessions.

There are mainly four methods to solve the subject and session differences.

1) Train subjects to show consistent and stable brain activities. This approach took months for the subjects to master.
2) Train emotion classifier in training trails and generalize to test trails. Sugiyama *et al.* [12] used reweighted samples in the source to train the target classifiers. This approach needs some training samples in the beginning of each session. The basic method was to pool samples from multiple recordings to train classifiers. However, the differences of the statistical distributions of subjects and sessions remain unsolved, making the effectiveness unstable.
3) Find some data structures that were invariant across subjects or sessions. For example, the common spatial patterns (CSPs) and sparse coding techniques project the EEG data to another space and transform new data to that space for classification. Krauledat *et al.* [13] extracted prototypical spatial filters with good generalization properties. Morioka *et al.* [14] proposed to learn a sparse feature representation with multiple subjects and used the resting-state activity of a new subject (target) as calibration to compensate for individual differences. These methods did not use unlabeled data in the target when they were available.
4) Combine the labeled training EEG data and the unlabeled test EEG data in the domain adaptation scheme and improve the target classification accuracy by using source supervisions.

This article solves the subject and sessions of differences with domain adaptation. Let $\mathcal{X} \in \mathbb{R}^{(n \times d) \times 1}$ be an EEG feature space, where n is the number of electrodes and d is the dimension of the feature extracted on each electrode. We denote $X \in \mathcal{X}$ as an EEG sample. In this article, the extracted feature is either power spectral density (PSD) [7] or differential entropy (DE) [8], and we concatenate features of all the electrodes to feed emotion recognition models. We denote $y \in \mathcal{Y}$ as the emotion label, where $\mathcal{Y} \in \mathbb{R}$ is the label space. Let $P(X)$ be the marginal probability distribution of X, then $\mathcal{D} = \{\mathcal{X}, P(X)\}$ is a domain [15]. $\mathcal{D}_S$ is labeled while $\mathcal{D}_T$ is unlabeled. They share the same feature space, i.e., $\mathcal{X}_S = \mathcal{X}_T$, but the marginal distributions are different, i.e., $P(X_S) \neq P(X_T)$. Most domain adaptation methods implicitly assume that the conditional distributions are the same, $P(Y_S|X_S) = P(Y_T|X_T)$, but there is no reason for this assumption to hold [16]. From the perspective of probability theory, here we divide domain adaptation methods into three categories.

The first domain adaptation method is marginal distribution adaptation (MDA). It is the most commonly used strategy that deals with the *covariate shift* problem. The MDA methods seek for a function ϕ to minimize $\|\phi(P(X_S)) - \phi(P(X_T))\|_{\text{dist}}$, where $\| \cdot \|_{\text{dist}}$ is some kind of distance measure. The Kullback–Leibler divergence [17] and the Jensen–Shannon divergence [18] are popular distances. In recent years, the maximum mean discrepancy (MMD) [19] and the multikernel MMD (MK-MMD) [20] have shown an advantage in domain adaptation. The most classical MDA method is the transductive component analysis (TCA) [21], which projects X_S and X_T to a reproducing kernel Hilbert space (RKHS) and solves a kernel to minimize the MMD between the mean values of the two domains in the low-dimensional space. Tzeng *et al.* [22] proposed to replace the kernel function with neural networks and minimized the MMD of the two domains at the last layer of DNNs, which was called the deep domain confusion (DDC). To reduce the reliance on the choice of the kernel function, Gretton *et al.* [20] proposed MK-MMD to replace MMD and offered efficient solutions. Long *et al.* [23] replaced MMD with MK-MMD in DDC, which is called the deep adaptation network (DAN). In addition, DDC only adapts the last layer, and DAN adapts the last three layers at the same time. Apart from the MMD-based methods, inspired by the generative adversarial networks (GANs) [24], the domain adversarial neural networks (DANNs) [25] and the adversarial discriminative domain adaptation (ADDA) [26] introduced adversarial training strategy to compel the shallow layers to learn domain-invariant representations and achieved the state of the art in the corresponding years. The MDA methods adapts only the marginal distributions, whereas the conditional information is not taken into consideration.

The second domain adaptation method is the conditional distribution adaptation (CDA). It looks for a function ϕ to minimize the distance between conditional distributions $\|\phi((Y_S|X_S)) - \phi((Y_T|X_T))\|_{\text{dist}}$. However, this task is usually hard since no label is available in the target [12]. In response, Wang *et al.* proposed the stratified transfer learning (STL) [27] to encourage intraclass transfer. STL begins with a rough classification in the target with the source classifier and reduces the intraclass MMD to tackle the variability in conditional distributions. An alternative to deal with this problem is not to explicitly infer the label of the target, but to enhance the conditional distribution similarity within a roundtrip starting from the source, passing the target, and returns to the source with source label consistency constrains. This idea stands for the associative domain adaptation (ADA) by Häusser *et al.* [28], which achieved state of the art on various benchmarks recently.

The third domain adaptation method is JDA. It seeks for a function ϕ to minimize $\|\phi(P(X_S, Y_S)) - \phi(P(X_T, Y_T))\|_{\text{dist}}$, where $P(X, Y) = P(X)P(Y|X)$. We use logarithm function to turn multiplication into addition. As logarithm is monotonically increasing, we could simultaneously conduct MDA and CDA as a proxy to JDA. Therefore, the goal changes to $\|\phi(P(X_S)) - \phi(P(X_T))\|_{\text{dist}} + \|\phi((Y_S|X_S)) - \phi((Y_T|X_T))\|_{\text{dist}}$. Long *et al.* proposed the first version of JDA [29], which adopted TCA to adapt the marginal distribution, and the CDA strategy is similar with STL. Recently, Rakotomamonjy *et al.* proposed the joint distribution optimal transportation (JDOT) to implement JDA. By integrating a label cost to the Kantorovich problem, JDOT solves the optimal classification on the target standing for a minimization of a bound on the target error [16].

This article proposes a novel JDA method that combines the advantage of MDA and CDA. The experimental results demonstrate the advantage of our method.

III. Methods

This section presents a novel JDA method by jointly adapting marginal distributions and conditional distributions. Different from the previous studies [10] which dispose EEG directly, we perform domain adaptation on their latent representations generated by neural networks, which may capture more domain-invariant structures. Moreover, in this article, the source labels are not only for training classifiers but also play a key role in reducing the conditional distribution differences.

A. Model Architecture

We draw an overview of the model architecture in Fig. 1, and then describe the algorithm mathematically. The method works with neural networks. The input is the EEG feature vector. If we feed the network with mini-batches, then during training, half of the batch is source samples, and the left comes from the target. The number of the two domain samples should keep basically the same. We divide the network into two parts, where θ_1 stands for the parameter set of the shallow layers, and θ_2 denotes the parameters of the deep layers. The network outputs the emotion label for each EEG sample. To make reliable predictions on the target, the updating guideline for θ_1 and θ_2 is not only to classify source emotion patterns but also to make the two domains more similar (adaptation). Considering that the shallower layers tend to generate task-invariant features, and the deeper layers are more likely to generate task-specific features [30], we use two distinct domain adaptation strategies on the whole network. In the shallower layers, θ_1 is not intently optimized to fit the classification task, but rather, the primary goal is to make the shallow representations of the two domains statistically similar, not class discriminative. This is achieved by adversarial training [24]. In the deeper layers, we update θ_2 to associate the two domains with source label consistency constraints [28], [31], while minimizing the source domain emotion classification error. The gradient also spreads to the shallow layers to affect θ_1, however, as gradient decreases along with the feedback layers, the main influence of the gradient is located in θ_2. The two strategies work together to endow the whole network with the ability to adapt to joint distributions.

Since the source is labeled, and the target is unlabeled, our goal is to classify target samples on the basis of the source supervision. During training, we concatenate the EEG features of both domains as mini-batches to feed the network, which uses the same set of parameters (connection weights) for both domains. Under θ_1, the shallow layers projects the data to a shallow latent space. The shallow representations then go into a domain predictor, a binary classifier who judges whether a shallow representation belongs to the source or the target. During feedforward, it acts as an ordinary classifier, whereas during the backward propagation, it compels the shallow layers to generate domain-invariant representations with gradient reversal, i.e., all gradients take the opposite numbers [25]. More specifically, the shallow layers are like the generator in a GAN, and the domain predictor is like the discriminator. They compete with each other in a zero-sum game. As the game reaches the Nash equilibrium, the shallow layers produce representations in which the domain predictor could not recognize the origin, and thus the marginal distributions are adapted. Except for getting into the domain predictor, the shallow representations also get into deep layers, where they are further processed to become deep representations followed by a label predictor. The label predictor outputs the emotion label. Before the softmax, the deep representations for the two domains are associated with source label consistency. The association reinforcement makes all previous layers, especially the deeper layers produce class discriminative representations. The conditional distributions are adapted here. During the test, the model inputs are the target samples only. The network predicts its domain labels and emotion labels.

B. Domain Adaptation Scheme

The method could be categorized with domain adaptation in neural networks, which makes the representations of the source and target as similar as possible (*assimilation*), while reducing the emotion classification error in the source as much as possible (*discrimination*) [22], [23], [28]. The training goal is formulated to minimize a loss function

$$\mathcal{L} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{similar}}. \tag{1}$$

The first item $\mathcal{L}_{\text{class}}$ encourages the emotion classification on the source, and the second item $\mathcal{L}_{\text{similar}}$ encourages the similarity of the representations of the two domains. As the source and target are made similar in terms of joint distributions, the model makes predictions on the target label with better reliability. In practice, we have

$$\mathcal{L}_{\text{class}} = H(\hat{y}, \mathrm{y}) = -\sum_{m=1}^{M} y_m \log \hat{y}_m \tag{2}$$

where H is the cross-entropy loss, y_m denotes the real emotion label, and $\hat{y}_m$ is the predicted emotion label. M is the number of emotion states to recognize. We are seeking the parameters θ_1 and θ_2 according to

$$\left(\widehat{\theta_1}, \widehat{\theta_2}\right) = \arg\min_{\theta_1,\theta_2} \mathcal{L}_{\text{class}}. \tag{3}$$

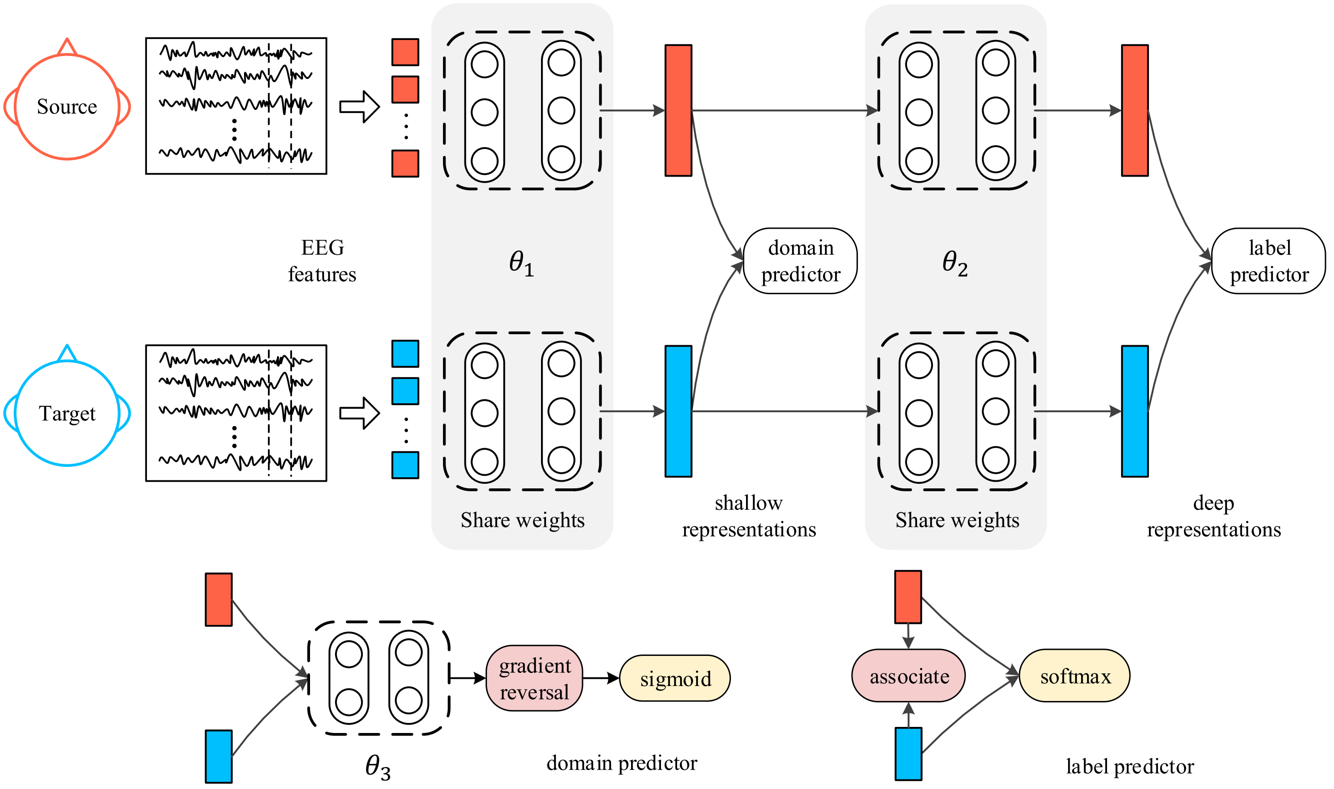

Fig. 1. Model architecture of the JDA. The source and the target share a same set of parameters. The shallower layers generate shallow representations, and we adapt the marginal distribution within a domain predictor with adversarial training. The deeper layers generate deep representations, and we adapt the conditional distribution within a (emotion) label predictor by association reinforcement. In this way, the joint distribution is approximately adapted.

In this article, we focus on $\mathcal{L}_{\text{similar}}$. It consists of two parts

$$\mathcal{L}_{\text{similar}} = \mathcal{L}_{\text{domain}} + \mathcal{L}_{\text{associate}}. \tag{4}$$

$\mathcal{L}_{\text{domain}}$ encourages the marginal distribution similarity, which is minimized with adversarial training. $\mathcal{L}_{\text{associate}}$ is for the conditional distribution similarity minimized with the association reinforcement. In the following parts, we show how to configure $\mathcal{L}_{\text{similar}}$ for network optimization.

C. Adversarial Domain Adaptation

We denote $x_i^s, i = 1, \ldots, n_s$ as EEG samples in $\mathcal{D}_S$, and $x_j^T, j = 1, \ldots, n_T$ are samples in $\mathcal{D}_T$. The shallower layers produce shallow representations for x_i^s and x_j^T. Inspired by the training scheme of GAN, we want the early layers (see Fig. 1) to generate adversarial samples to confuse the domain predictor in believing that the representations come from the same marginal distribution. We write the loss function as

$$\mathcal{L}_{\text{domain}} = H(\hat{p}, p) \tag{5}$$

where H is the cross-entropy loss, $p \in \{0, 1\}$ is the domain label, and $\hat{p}$ is the output of the domain classifier. If an EEG sample comes from the source, we have p= 0, and if it comes from the target, we have $p = 1$. We are seeking the parameters θ_1 and θ_3 according to

$$\widehat{\theta_1} = \arg\max_{\theta_1} \mathcal{L}_{\text{domain}} = \arg\min_{\theta_1} -\mathcal{L}_{\text{domain}} \tag{6}$$

and

$$\widehat{\theta_3} = \arg\min_{\theta_3} \mathcal{L}_{\text{domain}}. \tag{7}$$

As shown in Fig. 1, $\mathcal{L}_{\text{domain}}$ is computed at the end of the domain predictor. The derived gradients first update θ_3. To further update θ_1 according to (5), we reverse the gradients with gradient reversal module, which will be discussed in detail in the following parts.

D. Association Reinforcement

We denote A_i and B_j as the representations of x_i^s and x_j^T at the last layer of the label predictor (before softmax). The similarity between A_i and B_j is measured by the dot product $M_{ij} = \langle A_i, B_j \rangle$, which have high computational efficiency benefited from matrices multiplication.

The transition between the parts $\{\{A_i\}, \{B_j\}\}$ could be considered as a bipartite graph, which are more probable if the representations are more similar [28]. Therefore, like the softmax method, the transition probability from A_i to B_j is formulated as

$$P_{ij}^{ab} = P(B_j|A_i) := \frac{\exp(M_{ij})}{\sum_{j'} \exp(M_{ij})}. \tag{8}$$

If an imagery random walker starts from a labeled A_i to walk to an unlabeled B_j with probability P_{ij}^{ab}, and then returns back to another labeled A_j. The roundtrip probability for $A_i \rightarrow B_j \rightarrow A_j$ is written as a multiplication of P_{ij}^{ab} and P_{ij}^{ba}

$$P_{ij}^{aba} := \left(P^{ab}P^{ba}\right)_{ij}. \tag{9}$$

The association reinforcement is implemented by forcing the roundtrip to return to the same class (H: cross-entropy loss)

$$\mathcal{L}_{\text{walker}} = H\left(T, P^{aba}\right) \tag{10}$$

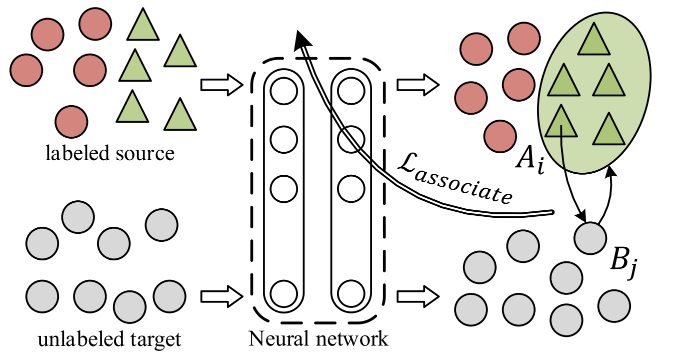

Fig. 2. Principle of association reinforcement. The neural network maps the labeled source and the unlabeled target to a latent space where a source sample A_i travels to a target sample B_j (with probability) and returns to arbitrary source samples holding the same emotion label with A_i. The network parameters are optimized to minimize $\mathcal{L}_{\text{associate}}$.

where

$$T_{ij} := \begin{cases} 1/|A_i|, & \text{class } (A_i) = \text{class}\left(A_j\right) \\ 0, & \text{else.} \end{cases} \tag{11}$$

In practice, the association process might neglect some hard-to-associate samples in the target. Therefore, a regularizer called $\mathcal{L}_{\text{visit}}$ is added to make sure that each target sample is visited with equal probability [28], which should be looser if the class distribution differences of the domains are considerable

$$\mathcal{L}_{\text{visit}} = H\left(V, P^{\text{visit}}\right) \tag{12}$$

where

$$P_j^{\text{visit}} := \sum_{x_i \in D_S} P_{ij}^{ab}, V_j := 1/\left|B_j\right|. \tag{13}$$

Therefore, the association of the two domains is enhanced with the combination of $\mathcal{L}_{\text{walker}}$ and $\mathcal{L}_{\text{visit}}$

$$\mathcal{L}_{\text{associate}} = \mathcal{L}_{\text{walker}} + \lambda \mathcal{L}_{\text{visit}}. \tag{14}$$

We are seeking the parameters θ_1 and θ_2 according to

$$\left(\widehat{\theta_1}, \widehat{\theta_2}\right) = \arg \min_{\theta_1, \theta_2} \mathcal{L}_{\text{associate}}. \tag{15}$$

Fig. 2 is an illustration of the association reinforcement. The neural network generates latent representations for both domains. The goal is to maximize the probability of any source samples to return to source samples holding the same emotion labels within roundtrips. There is no mandatory requirement to return each source sample to its original point. We perform association reinforcement in the label predictor shown in Fig. 1, which draws the two domains closer in terms of the conditional distributions with label consistency constraints.

E. Optimization With Backpropagation

In the optimization process, we use the following method to update the network parameters:

$$\theta_1 \leftarrow \theta_1 - \alpha \left(\frac{\partial \mathcal{L}_{\text{class}}}{\partial \theta_1} + \frac{\partial \mathcal{L}_{\text{associate}}}{\partial \theta_1} - \frac{\partial \mathcal{L}_{\text{domain}}}{\partial \theta_1} \right) \tag{16}$$

$$\theta_2 \leftarrow \theta_2 - \alpha \left(\frac{\partial \mathcal{L}_{\text{class}}}{\partial \theta_2} + \frac{\partial \mathcal{L}_{\text{associate}}}{\partial \theta_2} \right) \tag{17}$$

$$\theta_3 \leftarrow \theta_3 - \alpha \left(\frac{\partial \mathcal{L}_{\text{domain}}}{\partial \theta_3} \right) \tag{18}$$

where i is the sample index, and α is the learning rate.

During feedforward, the input samples are transformed with θ_1 to yield shallow representations followed by a logistic regression with parameter θ_3 to generate $\mathcal{L}_{\text{domain}}$. At the same time, the shallow representations are also fed into a subsequent transform with parameter θ_2, and the deep representations are thus obtained. The deep representations are used to compute $\mathcal{L}_{\text{associate}}$, and then used in the softmax to compute $\mathcal{L}_{\text{class}}$. The backpropagation follows (16)–(18). However, (16) is slightly different from the traditional gradient descent. To implement the adversarial training for θ_1 updating, we introduce a gradient reversal layer (GRL), where the gradients of $\mathcal{L}_{\text{domain}}$ handed back by the deep layers are reversed and multiplied with a parameter μ. It controls the tradeoff between the "generator" and the "discriminator."

IV. Experiments and Results

A. Data Description and Feature Extraction

We conduct experiments on two open emotion recognition data sets. The stimulus is both videos. Different from traditional static picture stimulus, such as the international affective picture system (IAPS), videos expose subjects to both visual and auditory stimulus, which can better evoke the corresponding emotions [8].

1) SEED [8], [32] contributed by Zheng *et al.* The emotion stimulus was the movie clips selected by 20 volunteers to make sure that they were coordinated in terms of valence and arousal levels. They assessed their emotions when watching movie clips from a materials pool using scores (1–5) and keywords (positive, neutral, and negative). Each clip lasted for about 4 min and elicited a single desired target emotion. Finally, to avoid ambiguity, only 15 film clips (five positive, five neutral, and five negative) were selected, which received a score of 3 or higher on the mean ratings from the 20 participants. During the EEG experiment, 15 healthy subjects (8 females, MEAN: 23.27, SD: 2.37) who passed the Eysenck personality questionnaire personality test watched the 15 clips, which are shown with the following procedure: a 5-s hint of start, the stimuli clip, a 45-s self-assessment, and a 15-s rest. The self-assessment guaranteed that the subjects showed the expected emotion. The details are in [8]. During movie-watching, 62-channel EEG was recorded with an ESI neuro scan system according to the international 10–20-system with 1000-Hz sampling rate. For preprocessing, the signals were downsampled to 200 Hz. Signals seriously contaminated by EMG and EOG are manually discarded. The signals then go through a bandpass filter between 0.3 and 50 Hz. After that, the EEG of each channel is divided into 1-s length segments without overlapping. The total number of samples for each subject were 3394, and the sample numbers for each of the three emotion states were basically the same. To facilitate the research on long-term models, each subject repeated the experiment for three times with an interval of about one week.

Each time is called a session. In general, the task on SEED is three-category valence classification, where both cross-subject and cross-session transfers are done.

2) DEAP [33] contributed by Koelstra *et al.* The emotion stimulus was 40 music videos selected by 14–16 volunteers via online self-assessment. Each video lasted for 1 min. During the experiment, 32 subjects watched the selected videos, and 32-channel EEG were recorded simultaneously with BioSemi EEG caps according to the international 10–20 systems. There were eight additional channels collecting peripheral physiological information, including galvanic skin response, skin temperature, blood volume pressure, respiration rate, electromyogram, and electrooculogram. Each subject rated the videos in terms of the levels of arousal, valence, like/dislike, dominance, and familiarity. More details about the experiment are available in [33]. For preprocessing, the signals were downsampled to 128 Hz and went through a bandpass filter between 4 Hz and 45 Hz. The EOG artifacts were removed. The total number of samples for each subject was 2520. Inconsistent with the previous study [32], we adopt the valence-arousal (VA) model, where each dimension has a value of ranging from 1 to 9. The coordinate system divides the VA space into four quadrants: 1) low-arousal–low-valence (LALV); 2) high-arousal–low-valence (HALV); 3) low-arousal–high-valence (LAHV); and 4) high–arousal-high-valence (HAHV). Considering that emotions often have fuzzy boundaries, and individual subjects have different rating scales, we add a gap to segment the quadrants to make sure that the EEG signals correspond to unambiguous target emotions [32]. More specifically, we discard the EEG samples whose valence and arousal ratings are both between 4.8 and 5.2. In general, the task on DEAP is the four category VA classification. We only do cross-subject transfer since each subject only participated in the experiment once.

Instead of using raw EEG time series, researches tended to extract features on some frequency bands and then conduct classification in EEG emotion studies [7], [8], [10]. There are five commonly used EEG frequency bands: 1) delta waves (1–3 Hz) associated with sleeping; 2) theta waves (4–7 Hz) with light meditation and sleeping; 3) alpha waves (8–13 Hz) with relaxation and happiness; and 4) beta waves (14–30 Hz) with consciousness and reasoning; and 5) gamma waves (31–50 Hz) with reasoning and other high-level information processing.

We employ two kinds of features to characterize the bands.

1) *PSD [7]:* It measures the spectral energy distribution for a time series on different bands, which has been widely used in EEG emotion recognition.
2) *DE [34]:* It is an extension of Shannon entropy measuring the complexity of a continuous random variable, which has been proved to be more accurate and stable than PSD in emotion recognition [35]. Existing researches have demonstrated that EEG signals on the five frequency bands approximately obey the Gaussian distribution [34], [35]. If a random variable obeys the Gaussian distribution, that is, $x \sim \mathcal{N}(\mu, \sigma^2)$, DE can simply be calculated by

$$h(x) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx$$
$$= \frac{\log 2\pi e \sigma^2}{2} \tag{19}$$

where σ is the variance of x, and e is the Euler constant.

For simplicity, we use DE to conduct experiments, and then repeat experiments using PSD to compare the results. The feature extraction details are as follows.

1) *SEED:* Each 1-s segment is decomposed with 256-point short-time Fourier transform using nonoverlapped Hanning window, and five frequency bands are thus obtained. PSD (DE) is computed on each of the five bands. PSD (DE) features form a 310-D vector since there are 62 channels. The features on SEED have been smoothed with the linear dynamic system (LDS) method. LDS is an effective approach to make use of the time dependency of emotion changes and filter out emotion-unrelated EEG components [32], [36].
2) *DEAP:* The feature extraction is similar with SEED. In the preprocessing, the slowest Delta waves were filtered out by the bandpass filter, therefore, we only compute the PSD (DE) on the latter four bands, which are faster brain waves. The PSD (DE) features form a 128-D vector since there are 32 channels. We applied moving average techniques to smooth the DEAP data. The window size is 5.

B. Transfer Learning Schemes

During training, the samples of the labeled source and the unlabeled target are used. During the test, the model deduces the emotion labels for the test samples. It is a transductive machine learning design. Considering the different application scenarios, we define two kinds of transfer schemes.

1) *Many-to-One Transfer (Denoted as $\mathcal{M} \rightarrow \mathcal{O}$):* In the cross-subject transfer, we spare one subject as the target and put the data of all the rest subjects together as the source. In the cross-session transfer (on SEED only), the target is one-session EEG of one subject, and the source is the combination of his/her EEG from other sessions. The $\mathcal{M} \rightarrow \mathcal{O}$ scheme makes sense when multiple existing subjects (sessions) are available.
2) *One-to-One Transfer (Denoted as $\mathcal{O} \rightarrow \mathcal{O}$):* In the cross-subject transfer, the target is one subject's EEG, and the source is another subjects' EEG. In the cross-session transfer, the target is the one subject's one-session EEG, and the other sessions take a turn as the source. The $\mathcal{O} \rightarrow \mathcal{O}$ scheme makes sense when only one existing subject (session) is available.

C. Cross-Subject Transfer

For methods, such as TCA, the computation was heavy in $\mathcal{M} \rightarrow \mathcal{O}$ if all the source samples are included. Therefore, Zheng and Lu [10] randomly selected a subset of 5000 samples to form the source. In our experiment, we compare the performance of using 5000 samples and using all available samples in the source.

We use multilayer perceptron (MLP) to implement the DNN. After a series of experiments and observations, we fix the model structure for SEED as 310 (input layer)-128 (middle layer)-3 (output layer). The 128-D embedding is the shallow representation, and the 3-D embedding is deep representation. The deep embedding is followed by the SoftMax classification. Therefore, the model is a three-layer full-connection neural network, which could approximate complex functions with sufficient neurons. The activations are all ReLU. We have explored more layers to generate the representations, for example, a 310-256-128 (shallow representation)-32-3 (deep representation) model to conduct the experiment, where the training time was longer, but the accuracy improvement was not significant. Therefore, we use the three-layer network to conduct a further experiments, which is easy to train and produces reliable representations. The model on DEAP has more layers. The network was 128-64-64 (shallow representation)-16-2 (deep representation). The structures remain unchanged in both cross-subject and cross-session situation.

For both structures, the shallow representations are fed into a domain predictor, which has two neurons performing SoftMax classification. Adam shows better performance than SGD, RMSProp, and AdaGrad. The learning rate is 1e-3. We used the default parameters in Adam: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\in = 10e - 8$. The batch size is 96. The parameter λ is searched between 0.2 and 1 at a pace of 0.2, and we find that 0.6 is suitable in both DEAP and SEED. The parameter λ is set as 0.6 for all the experiments. To avoid overfitting, we use $L2$ regularizes (1e-3) in the networks for SEED. The training epoch is set to 8 K with early stop (patience = 10) applied on the training loss. We do the experiments with TensorFlow libraries with an NVIDIA GeForce GTX 1080 GPU.

Table I shows the mean accuracies for $\mathcal{M} \rightarrow \mathcal{O}$ transfer. On SEED, to reduce the evaluation complexity, we used one session's data for each subject to conduct transfer, which is inconsistent with [10]. We used the neural network model to evaluate the training accuracy of each session and selected the session holding the best performance for the experiment. Source only (SO) means training on the source, and test on the target. It is a baseline accounting for "nontransfer." Target only (TO) means training and test on the target. Kernel-PCA (KPCA) [37], TCA, transductive SVM (TSVM) [38], and transductive parameter transfer (TPT) [39] are four methods used in [10]. They stand for conventional transfer learning methods. DANN uses an adversarial training strategy to close the two marginal distributions in the shallow representations generated by the shallow layer, i.e., $\mathcal{L}_{\text{similar}} = \mathcal{L}_{\text{domain}}$. MMD replaces $\mathcal{L}_{\text{domain}}$ with MMD distance in (4) to adapt the marginal distributions, i.e., $\mathcal{L}_{similarity} = \mathcal{L}_{\text{MMD}}$. ADA uses association reinforcement for the CDA, i.e., $\mathcal{L}_{similarity} = \mathcal{L}_{\text{associate}}$. Our method jointly adapts the marginal and the conditional distributions, i.e., $\mathcal{L}_{\text{similar}} = \mathcal{L}_{\text{domain}} + \mathcal{L}_{\text{associate}}$. DANN and ADA can be seen as ablation tests to our method.

From TO to SO, the mean accuracies for SEED and DEAP drops for about 38% and 25%, respectively. This is due to the domain differences. Among the four conventional methods, TPT (76.31%) outperforms the others on SEED. The performance of DANN and MMD are basically equal, and both are better than the conventional methods. ADA shows the advantage in emotion recognition, where the accuracies reach 84.47% (SEED) and 60.38% (DEAP) using 5000 source samples. Our method beats ADA by around 2% on SEED and DEAP when 5000 samples are used. When all the source samples are used, the accuracies on SEED improves for about 2% to 3%. One DEAP, the improvement is not visible. The main finding remains and our method yields the highest accuracy.

TABLE I
CROSS-SUBJECT MANY-TO-ONE TRANSFER

Method	SEED	DEAP
SO	58.23 (18.56)	41.38 (17.22)
TO	96.21 (2.22)	66.03 (10.65)
KPCA [37]	61.28 (14.62)	46.75 (14.34)
TCA [21]	63.64 (14.88)	50.28 (15.70)
TSVM [38]	72.53 (14.00)	54.34 (15.38)
TPT [39]	76.31 (15.89)	55.25 (14.23)
DANN [25]	78.76 (7.33)	55.95 (9.97)
DANN-all	81.65 (9.92)	56.14 (10.60)
MMD [19]	78.93 (9.56)	55.33 (9.82)
MMD-all	80.88 (10.1)	56.76 (12.12)
ADA [28]	84.47 (10.65)	60.38 (10.30)
ADA-all	86.89 (8.99)	60.35 (10.33)
Ours	**86.70 (11.39)**	**62.54 (11.20)**
Ours-all	**88.28 (11.44)**	**62.66 (10.45)**

The results are shown in MEAN (STD). Best results in bold. 5000 samples in the source are used. 'All' means using all the samples in the source are used.

TABLE II
CROSS-SUBJECT ONE-TO-ONE TRANSFER

Method	SEED	DEAP
SO	57.50 (16.6)	40.27 (18.89)
TO	96.21 (2.22)	66.03 (10.65)
DANN [25]	61.57 (14.01)	50.50 (9.79)
MMD [19]	59.15 (12.28)	48.87 (12.39)
ADA [28]	63.19 (14.67)	55.91 (13.08)
Ours	**65.94 (15.99)**	**59.39 (15.55)**

We also notice that the standard deviations of SO are high, indicating that the performance varies a lot across subjects. As a comparison, the deviations of TO are small, indicating the consistency of EEG data for a same subject in one session.

Table II shows the mean accuracies of $\mathcal{O} \rightarrow \mathcal{O}$ transfer. The accuracy down cliffs also emerge from TO to SO for about 39% on SEED and 26% on DEAP. DANN and MMD restore the accuracies to around 60% on SEED, and 50% on DEAP, respectively. ADA beats all the comparing methods, and our method shows the advantage of both data sets.

D. Cross-Session Transfer

Table III shows the mean accuracies for cross-session transfer. The left column contains the $\mathcal{M} \rightarrow \mathcal{O}$ results. The

TABLE III
CROSS-SESSION TRANSFER RESULTS

Method	$\mathcal{M} \to \mathcal{O}$	$\mathcal{O} \to \mathcal{O}$
SO	81.28 (11.74)	76.91 (13.31)
TO	97.69 (1.38)	97.69 (1.38)
DANN [25]	83.15 (12.01)	78.86 (13.02)
MMD [19]	84.38 (12.05)	78.34 (12.64)
ADA [28]	89.13 (7.13)	85.29 (9.57)
Ours	**91.17 (8.11)**	**87.99 (8.78)**

The experiments are conducted on SEED.

gap between SO and TO (around 16%) is not as large as that of the cross-subject case, indicating the EEG for a single subject show some consistency over time. DANN and MMD improve the accuracy by 2%–3%. As a comparison, ADA improves for around 8%. Our method reaches an accuracy of 91.17%. The right column shows $\mathcal{O} \to \mathcal{O}$ results. Our method improves the mean accuracy from SO by 11%, yielding the best performance comparing with the other methods. However, the results are inferior to $\mathcal{M} \to \mathcal{O}$. Comparing with the results in Table I, we conclude that the cross-session results are better than all the cross-subject results. Therefore, transferring knowledge within a same subject is more favorable.

E. Comparison of EEG Features

To evaluate the impact of the feature on the results, we repeat the experiments using PSD. The experimental settings and the model implementations stay the same. The results of our method are summarized in Fig. 3. For cross-subject $\mathcal{M} \to \mathcal{O}$ transfer, the performance of DE beats PSD by 7.5% and 3.5% on SEED and DEAP, respectively. For cross-subject $\mathcal{O} \to \mathcal{O}$ transfer, DE shows an advantage for about 3.6% and 2.4% on SEED and DEAP, respectively. DE is superior to PSD in cross-subject domain adaptation. For cross-session $\mathcal{M} \to \mathcal{O}$ transfer, DE has a 6.8% advantage over PSD on SEED. In the cross-session $\mathcal{O} \to \mathcal{O}$ transfer, DE is better than PSD for around 4.2%. DE beats PSD remarkably in a cross-session case. In general, DE is more favorable than PSD in EEG emotion recognition, as well as the cross-subject and cross-session domain adaptation. Our transfer learning algorithm improves the accuracy for a considerable levels using PSD, which shows the versatility of our method.

F. Visualizations

To show the effectiveness of our method in an intuitive way, we project the latent representations on 2-D planes with t-SNE techniques. T-SNE is a nonlinear dimension-reduction method, which maintains data structures in low-dimensional spaces [40]. For simplicity, we randomly select one target in cross-subject $\mathcal{M} \to \mathcal{O}$ transfer on SEED and draw the TSNE embedding on 2-D planes. Fig. 4 shows what happened in the neural network. There are some notable phenomena.

1) The neural and negative emotions are not easy to separate, while positive emotions are easy to recognize. It is inconsistent with the previous studies [11].
2) Comparing with SO, DANN makes the source and the target statistically similar to the adversarial training. Although the evidence is not clearly visible, the effects of DANN are revealed in the classification accuracies. The representations of DANN are lack of visible class distinguishability because only the marginal distributions are adapted.
3) Under the influence of association reinforcement, ADA adapts the conditional distributions, making the samples fall into clusters according to emotion labels. However, the margins between neutral and negative emotions are not very clear.
4) Comparing with DANN, our method generates more label distinguishable representations. Comparing with ADA, our method generates more compact (concentrated) representations. In terms of classification accuracy, our method beats DANN by around 8%, while the advantage over ADA is relatively weak (around 2%). Taken together, CDA is more powerful than MDA. More specifically, in (4), $\mathcal{L}_{\text{associate}}$ plays a major role, and $\mathcal{L}_{\text{domain}}$ plays a supporting role.

V. DISCUSSION

Supervised learning is the conventional way to acquire EEG emotion recognition models. However, in some cases, the label information is unavailable, and the labeling work is expensive and time consuming, which leads us to find effective ways to borrow useful knowledge from other subjects or existing sessions to the present use. We develop domain adaptation algorithms to achieve this goal. The experimental results offer some inspirations for further researches.

1) The EEG of individual subject changes over time, which means the model becomes obsolete over time. We find the intrasession differences of a same subject are smaller than the intrasubject differences for a considerable level. The reasons are as follows: the cross-session transfer shows an advantage for 4.5% ($\mathcal{M} \to \mathcal{O}$) and 22% ($\mathcal{O} \to \mathcal{O}$) over cross-subject transfer on SEED. Therefore, if conditions permit, the cross-session transfer is more favorable than the cross-subject transfer.
2) In the cross-subject transfer, $\mathcal{M} \to \mathcal{O}$ beats $\mathcal{O} \to \mathcal{O}$ by about 20% on SEED. On DEAP, the advantage shrink to 3%. In general, $\mathcal{M} \to \mathcal{O}$ transfer is better than $\mathcal{O} \to \mathcal{O}$ transfer. In the cross-session transfer, the advantage of $\mathcal{M} \to \mathcal{O}$ is around 3%. The benefit of $\mathcal{M} \to \mathcal{O}$ is that the source is composed of samples from multiple subjects with distribution differences. As the diversity of the labeled samples is abundant in the feature space, the generalization ability of the models improve.
3) DE feature is better than the PSD feature in our task. DE shows a constant advantage over PSD in all the experiments.
4) Conventional MDA focus on the marginal distributions only and the conditional information is often neglected. In this article, we design an optimization scheme to endow the network with the ability to approximately

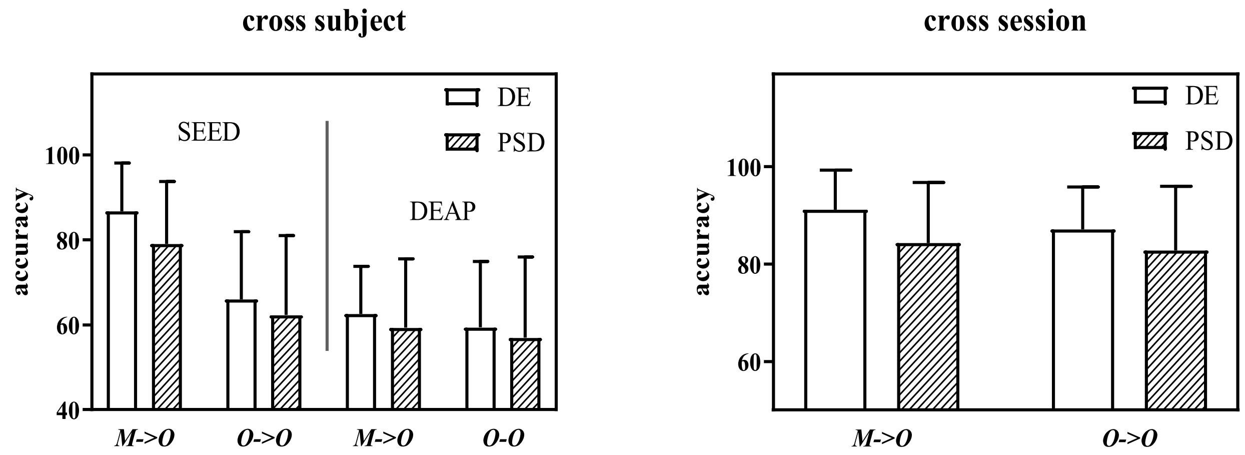

Fig. 3. Performance comparison of DE and PSD feature for both $\mathcal{M} \rightarrow \mathcal{O}$ and $\mathcal{O} \rightarrow \mathcal{O}$ schemes. Left: cross-subject results on SEED and DEAP. Right: cross-session results on SEED. In all the experiments, DE is better than PSD in terms of the mean accuracy.

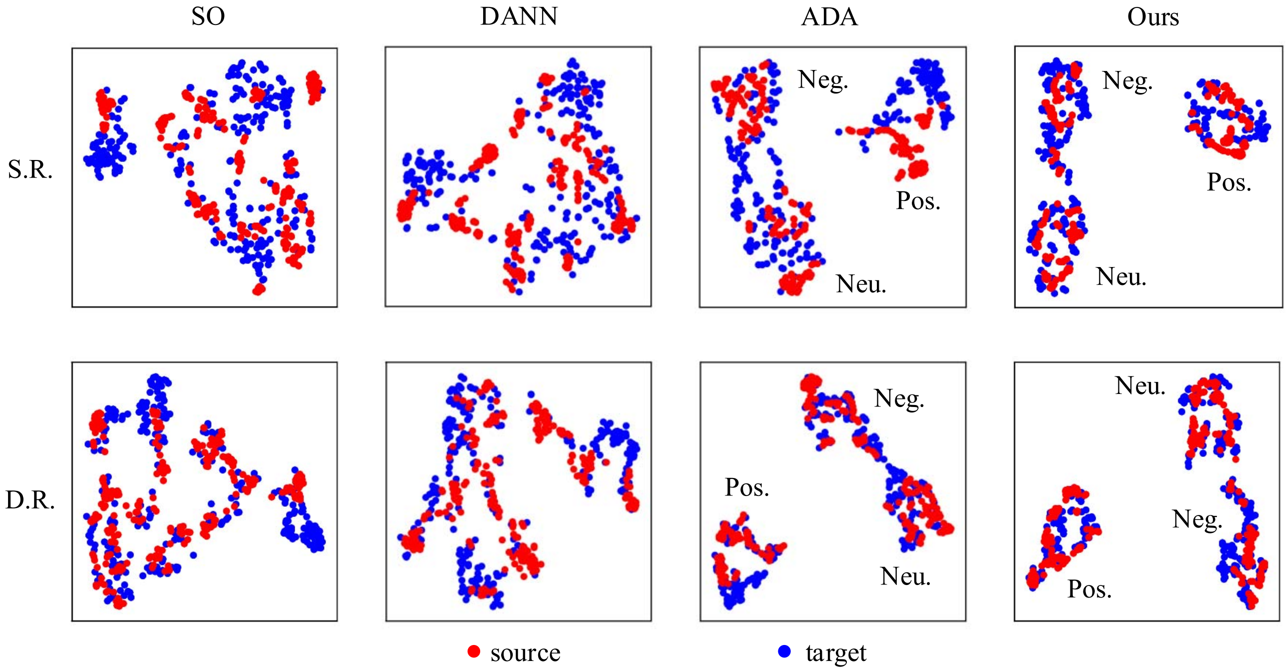

Fig. 4. Visualizations of the latent representations with the TSNE technique. For simplicity, we only take SEED as an illustration. The top row is for the shallow representations (S.R.), and the bottom row is for the deep representations (D.R.). Each column stands for a domain adaptation method. The S.R. and D.R. of SO and DANN do not gather into three distinct groups according to emotion labels (Pos: positive, Neu: neutral, Neg: negative), because only the marginal distributions are adapted. By taking the conditional distribution into consideration, ADA and our method generate visible emotion clusters. Finally, comparing with ADA, our method makes the clusters more compact by jointly adapting the marginal and conditional distributions.

adapted to the joint distribution. We have demonstrated that the cooperation of MDA and CDA could act as a proxy for JDA. Considering the different functions of MDA and CDA, we build the relationship between different network layers with different adaptation strategies, i.e., the shallow layers mainly conduct MDA, and the deep layers conduct CDA. In this way, our method not only makes the distributions statistically similar but also class sensitive. The results show that the proposed method is better than the existing MDA and CDA approaches. JDA should be a valuable future research direction in domain adaptation.

5) The method does not require complex model structures. After exploring different depths, we find that the results remain with simple neural networks. In practice, it means the training burden is light. In addition, the method can readily add to the existing models. We notice that the $L2$ penalty on the loss function is useful to improve the model generalization.

Concerning future works, there are two possible directions.

1) *JDA in Deep Convolutional Neural Networks (CNNs):* In our previous work [42], we organized the multichannel EEG as topology-preserving 2-D maps and used CNN to excavate the local and global correlations among electrodes. The results showed the potential of CNN in emotion recognition. We will evaluate the JDA method in CNN, and find suitable correspondence between domain adaptation strategies and different network layers. The

results could provide valuable references for future researches.

2) *Fast Online JDA Algorithms:* Our method, together with most of the existing methods are offline, which means the source and target samples are all available before knowledge transfer. However, the application scenarios may be online, where the source or the target, or both of them are collected in stream. Therefore, online JDA algorithm is an important research issue.

VI. Conclusion

We have proposed a novel domain adaptation method for EEG emotion recognition, which shows superiority in both cross-subject and cross-session adaptation. It integrates MDA (task-invariant features) and CDA (task-specific features) in a unified framework and needs no label information in the target domain to accomplish JDA. We compare it with a series of conventional and recent transfer learning algorithms. The results demonstrate that the method significantly outperforms the other approaches in terms of accuracies. The visualization analysis offers an insight into the influence of JDA on the representations. This article facilitates and promotes the applications of EEG emotion recognition in practice.

References

[1] H. Gunes, B. W. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops (FG)*, 2011, pp. 827–834.

[2] R. Cowie *et al.*, "Emotion recognition in human–computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[3] S. Kuusikko *et al.*, "Emotion recognition in children and adolescents with autism spectrum disorders," *J. Autism Develop. Disorders*, vol. 39, no. 6, pp. 938–945, 2009.

[4] J. Joshi *et al.*, "Multimodal assistive technologies for depression diagnosis and monitoring," *J. Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[5] G. L. Ahern and G. E. Schwartz, "Differential lateralization for positive and negative emotion in the human brain: EEG spectral analysis," *Neuropsychologia*, vol. 23, no. 6, pp. 745–755, 1985.

[6] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul./Sep. 2019.

[7] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul.–Sep. 2014.

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

[9] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup, "Transfer learning in brain–computer interfaces," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 20–31, Jan. 2016.

[10] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2732–2738.

[11] J. Li, Z. Zhang, and H. He, "Hierarchical convolutional neural networks for EEG-based emotion recognition," *Cogn. Comput.*, vol. 10, no. 2, pp. 368–380, 2018.

[12] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *J. Mach. Learn. Res.*, vol. 8, pp. 985–1005, May 2007.

[13] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards zero training for brain–computer interfacing," *PLoS ONE*, vol. 3, no. 8, 2008, Art. no. e2967.

[14] H. Morioka *et al.*, "Learning a common dictionary for subject-transfer decoding with resting calibration," *NeuroImage*, vol. 111, pp. 167–178, May 2015.

[15] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[16] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3730–3739.

[17] J. M. Joyce, "Kullback–Leibler divergence," in *International Encyclopedia of Statistical Science*. Heidelberg, Germany: Springer, 2011, pp. 720–722.

[18] B. Fuglede and F. Topsoe, "Jensen–Shannon divergence and Hilbert space embedding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2004, p. 31.

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

[20] A. Gretton *et al.*, "Optimal kernel choice for large-scale two-sample tests," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1214–1222.

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

[22] E. Tzeng *et al.*, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint, arXiv:1412.3474*, 2014.

[23] M. Long *et al.*, "Learning transferable features with deep adaptation networks," *arXiv preprint, arXiv:1502.02791*, 2015.

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

[25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint, arXiv:1511.06434*, 2015.

[26] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2017, pp. 2962–2971.

[27] J. Wang *et al.*, "Stratified transfer learning for cross-domain activity recognition," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, 2018, pp. 1–10.

[28] P. Häusser, T. Frerix, A. Mordvintsev, and D. Cremers, "Associative domain adaptation," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, vol. 2, 2017, pp. 2784–2792.

[29] B. Wang, W. Li, W. Fan, X. Chen, and D. Wu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2959–2963.

[30] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[31] P. Haeusser, A. Mordvintsev, and D. Cremers, "Learning by association—A versatile semi-supervised training method for neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 626–635.

[32] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul.–Sep. 2019.

[33] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[34] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *Proc. IEEE 35th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, 2013, pp. 6627–6630.

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

[36] L. C. Shi and B. L. Lu, "Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, 2010, pp. 6587–6590.

[37] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 536–542.

[38] R. Collobert, F. H. Sinz, J. Weston, and L. Bottou, "Large scale transductive SVMs," *J. Mach. Learn. Res.*, vol. 7, pp. 1687–1712, Aug. 2006.

[39] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 357–366.

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