Joint Feature Adaptation and Graph Adaptive Label Propagation for Cross-Subject Emotion Recognition from EEG Signals

Yong Peng, *Member, IEEE,* Wenjuan Wang, Wanzeng Kong, *Member, IEEE,* Feiping Nie, *Senior Member, IEEE,* Bao-Liang Lu, *Fellow, IEEE,* and Andrzej Cichocki, *Fellow, IEEE*

Abstract—Though Electroencephalogram (EEG) could objectively reflect emotional states of our human beings, its weak, non-stationary, and low signal-to-noise properties easily cause the individual differences. To enhance the universality of affective brain-computer interface systems, transfer learning has been widely used to alleviate the data distribution discrepancies among subjects. However, most of existing approaches focused mainly on the domain-invariant feature learning, which is not unified together with the recognition process. In this paper, we propose a joint feature adaptation and graph adaptive label propagation model (JAGP) for cross-subject emotion recognition from EEG signals, which seamlessly unifies the three components of domain-invariant feature learning, emotional state estimation and optimal graph learning together into a single objective. We conduct extensive experiments on two benchmark SEED_IV and SEED_V data sets and the results reveal that 1) the recognition performance is greatly improved, indicating the effectiveness of the triple unification mode; 2) the emotion metric of EEG samples are gradually optimized during model training, showing the necessity of optimal graph learning, and 3) the projection matrix-induced feature importance is obtained based on which the critical frequency bands and brain regions corresponding to subject-invariant features can be automatically identified, demonstrating the superiority of the learned shared subspace.

Index Terms—Electroencephalogram (EEG), emotion recognition, feature adaptation, graph learning, label propagation.

1 Introduction

EMOTION is essential in interpersonal communication, and influences our rational decision-making, cognition and behavior expression, which makes emotion recognition become a research hotspot in diverse fields such as information science, cognitive science and biomedical engineering. Since emotion usually refers to a state of mind that occurs spontaneously rather than consciously and is often accompanied by physiological reactions in central nervous and periphery, EEG generated from electrical activities in human cerebral cortex offers us a more reliable data source for objective emotion recognition than other data modalities such as facial expression, speech, text and peripheral physiological signals [1], [2]. With the fast development of wearable signal acquisition devices and EEG data processing techniques, EEG-based emotion recognition has attracted increasing attention in both academia and industry.

The typical paradigm of affective brain-computer interface systems (aBCIs) to evoke emotional EEG data includes four stages, *i.e.*, EEG signal acquisition, signal preprocessing, feature extraction and recognition [3]. In this paper, we focus our attention on the last two stages which are closely related to machine learning the most. Though existing machine learning models made lots of efforts on EEG-based emotion recognition, most of them followed the assumption that training and test EEG samples are independent and identically distributed (i.i.d.) [4], [5]. This might be not the case in real applications because the non-stationary property of EEG data easily causes the distribution divergences among different subjects. Then, the generalization performance of conventional machine learning models cannot be guaranteed since the underlying i.i.d. assumption on training and test data is no longer valid. To this end, transfer learning has been widely used to minimize the discrepancies for EEG data collected from different subjects.

Basically, transfer learning aims to utilize the knowledge from one domain (*i.e.*, source domain) to facilitate the learning task in the other domain (*i.e.*, target domain). In EEG-based emotion recognition, the source domain usually consists of EEG samples from one/more subject(s) which are fully labeled and the target domain is composed of EEG samples from another subject which are fully unlabeled. This is corresponding to the typical transfer learning paradigm, unsupervised domain adaptation, which is mainly achieved by projecting both source and target data into a shared subspace to minimize the domain discrepancies. Then, emotion recognition is performed on aligned EEG data. That is, most of existing studies completed the EEG-based cross-subject emotion recognition in a two-stage manner of performing the domain invariant feature learning

- *Yong Peng, Wenjuan Wang and Wanzeng Kong are with School of Computer Science and Technology, Hangzhou Dianzi University, and Zhejiang Key Laboratory of Brain-Machine Collaborative Intelligence, Hangzhou 310018, China. (Corresponding author: Yong Peng, yongpeng@hdu.edu.cn; Wanzeng Kong, kongwanzeng@hdu.edu.cn)*
- *Feiping Nie is with School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China.*
- *Bao-Liang Lu is with Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.*
- *Andrej Cichocki is with Center for Computational and Data-Intensive Science and Engineering, Skolkov Institute of Science and Technology, Moscow 143026, Russia.*

first and then the emotional state estimation [6], [7]. This inevitably causes the sub-optimality which might degenerate the recognition performance. Besides the recognition accuracy, most existing studies only visualized the aligned distributions of source and target EEG data and did not sufficiently investigate the properties of the learned shared subspace in emotion expression [8], [9], [10].

To solve the above mentioned limitations, we propose a joint feature alignment and graph adaptive label propagation model (JAGP) for EEG-based cross-subject emotion recognition. Compared with the existing studies, the present work has the following contributions.

- JAGP is a unified framework for joint EEG data distribution alignment and graph-based semi-supervised emotional state estimation of target samples. On one hand, the estimated emotional states serve as soft labels for better aligning the conditional distributions of source and target EEG data; on the other hand, better aligned data can better facilitate the emotional state estimation of target data as the number of model optimization iterations increases.
- Usually, the graph is first built from the unaligned source and target data, which is then kept fixed during the model learning process; therefore, the semi-supervised emotional state propagation on such graph is obviously not accurate enough to characterize the common information between source and target data. In JAGP, the graph is dynamically updated based on the source and target data in their respective shared subspace representations, which is more competent for target emotional state estimation.
- In this work, we are not only concerned with the cross-subject emotion recognition performance by using the shared subspace representations of source and target data, but also provide sufficient analysis on the properties of the shared subspace. From the perspective of transfer learning, it mainly contributes to preserve the common information and rule out the inconsistent information across subjects. Therefore, we can quantitatively calculate the feature importance and further obtain the common activation patterns of critical EEG frequency bands and channels across subjects.

The remainder of this paper is organized as follows. In section 2, we briefly review EEG-based emotion recognition and transfer learning in aBCIs. In section 3, we present the model formulation and optimization algorithm of JAGP in detail. Experiments are conducted and the results are analyzed in section 4. Some discussions on the proposed JAGP model are provided in Section 5. Section 6 concludes the whole paper.

Notations. In this paper, model variables and parameters are represented by Greek alphabets such as α, λ, γ, δ, and η. The EEG frequency bands are represented by *Delta*, *Theta*, *Alpha*, *Beta*, and *Gamma*. Matrices and vectors are respectively denoted by boldface uppercase and lowercase letters. For matrix $\mathbf{A}$, $\mathbf{a}^i$ denotes its i-th row and $\mathbf{a}_j$ denotes its j-th column. a_{ij} denotes its (i,j)-th element. $\mathbf{1}$ represents an all-one matrix or vector whose size can be inferred from context.

2 Related Works

2.1 EEG-based Emotion Recognition

Generally, the existing EEG features for emotion recognition can be divided into four types, *i.e.*, time domain features, frequency domain features, time-frequency domain features and spatial information features [11]. Jenke *et al.* reviewed the popular EEG features and feature selection methods in emotion recognition [12]. Currently, researchers rely more on machine learning models for EEG feature transformation and state recognition. In [13], Zheng *et al.* provided a comprehensive study on the performance evaluation of different EEG features and machine learning models in emotion recognition. Moreover, EEG features from different frequency bands were tested and they found that the *Gamma* band is the most important one in emotion recognition. In [5], an adaptive feature importance learning model was proposed to automatically identify critical EEG frequency bands and channels in emotion recognition.

Since deep learning models usually have powerful nonlinear feature representation capability than shallow ones, they appear increasingly frequently in the field of EEG-based emotion recognition. A fusion model of graph convolutional neural network and long-short memories neural networks for EEG-based emotion classification was proposed in [14]. In [15], a dynamic graph convolutional neural network (DGCNN) was proposed for EEG emotion recognition in which the graph modeled the multichannel EEG features. As an extension of DGCNN, sparseDGCNN additionally imposes a sparseness constraint on the graph inspired by the consensus that different brain regions sampled by EEG electrodes may be related differently to brain functions [16]. Deep neural networks can sometimes perform end-to-end learning by taking raw EEG data as input and directly outputting recognition results [17]. Though deep learning models generally obtained better performance, they usually cost more time in model training and also more training samples to fit the large parameter spaces. Besides, their results are usually difficult to interpret due to the black-box training mode [18], [19]. For example, based on the weight distributions of trained deep belief networks, the critical EEG frequency bands for cross-session emotion recognition were identified [20]. However, its underlying mechanism is still not intuitive enough. Recent advances in EEG-based emotion recognition can be found in [3], [21], [22].

2.2 Transfer Learning in EEG Emotion Recognition

EEG data is weak and non-stationary and therefore distribution discrepancies often appear in EEG data collected from different subjects. Therefore, the i.i.d assumption is usually violated if we train a machine learning model on data collected from one subject and then test it on data from another subject. To this end, transfer learning [23], [24] has been widely used to minimize the divergence between cross-subject EEG data, for the purpose of enhancing the universality of aBCIs.

Generally, existing transfer learning models in EEG-based BCIs can be divided into linear and nonlinear ones. In [10], a dynamic distribution alignment with dual-subspace mapping method (DDADSM) was proposed for cross-subject driver mental state detection. The two merits

of DDADSM are two folds; on one aspect, two optimal subspaces were explored to align the source and target distributions and on the other aspect, adaptive weights between the marginal and conditional distributions were learned. A two-stage transfer model termed NMF-TL was proposed to first use the non-negative matrix factorization to learn a common subspace shared by training and test EEG data, and then form an augmented data set by combining the original data representations and shared subspace representations for classification [25]. In [26], a multi-source domain transfer discriminative dictionary model was proposed for EEG-based emotion recognition by integrating transfer learning and dictionary learning into a single model objective. To differentiating the contributions of different sources, the multisource style transfer mapping (MS-STM) was proposed by first performing source selection and then reducing the EEG differences between the target and each source by style transfer mapping [27], [28].

By mapping data from the original representation into the reproducing kernel Hilbert space (RKHS), some kernel trick based transfer models were applied to model the nonlinear properties of EEG data. In [6], two cross-subject transfer models, *i.e.*, transfer component analysis (TCA) and kernel principal component analysis (KPCA), were used for building personalized EEG-based affective models. Specifically, TCA aims at mitigating the distribution mismatch by minimizing the maximum mean discrepancy (MMD) in the RKHS [29] and KPCA is the nonlinear extension of PCA by kernel trick. Joint distribution adaptation (JDA) jointly adapt both the marginal and conditional distributions of source and target data in a principled dimensionality reduction procedure [30]. The geodesic flow kernel (GFK) domain adaptation constructs a geodesic flow to link the subspaces of the source and target domains by an infinite number of interpolating subspaces in-between [8]. The maximum independence domain adaptation (MIDA) model maximizes the independence between the projected samples and their respective domain features according to the Hilbert-Schmidt independence criterion (HSIC) [31], [32].

Besides the kernel-based nonlinear transfer models, neural networks are more often used in subject-independent EEG emotion recognition due to their powerful nonlinear feature learning ability. In [9], deep networks were used to simultaneously minimize the recognition error on source data and force the latent representation similarity (LRS) of source and target data. To reduce the risk of negative transfer, based on the local and global attention mechanism, a transferrable attention neural network was proposed to learn the emotional discriminative information by highlighting the transferrable brain regions data and samples adaptively [33]. According to the emotional brain's asymmetries between left and right hemispheres, EEG samples of both hemispheres are separately mapped into discriminative feature spaces [34], [35]. Moreover, a global and two local domain discriminators were used to suppress the distribution shift between the training and test data. Wu *et al.* made a thorough review on the transfer learning models in EEG-based BCIs [36], [37].

Though transfer learning has been widely used in EEG-based emotion recognition to align the EEG data from different subjects [38], most of existing studies put the emphasis on domain-invariant feature learning [8], [9], [10]. Because the two stages of feature alignment and emotional state recognition are not effectively unified together, this sequential execution manner inevitably causes the sub-optimality. In [39], a graph adaptive knowledge transfer (GAKT) model was developed to jointly optimize the target labels and the domain-free features. However, the graph to perform label propagation in GAKT is directly built from the unaligned source and target data, which is kept fixed during the whole model learning process. It is likely to mislead the label propagation as the number of iterations increases.

3 The Proposed Model

In this section, we first formulate the model objective of JAGP and then describe its optimization algorithm.

3.1 Model Formulation

The basic settings of cross-subject EEG emotion recognition are stated as follows. EEG samples from source subject(s) are fully labeled, *i.e.*, $\{\mathbf{X}_s \in \mathbb{R}^{d\times n_s}, \mathbf{Y}_s \in \mathbb{R}^{n_s\times c}\} \triangleq \{(\mathbf{x}_1,\mathbf{y}^1),\cdots,(\mathbf{x}_{n_s},\mathbf{y}^{n_s})\}$, and EEG samples from target subject are totally unlabeled, *i.e.*, $\mathbf{X}_t \in \mathbb{R}^{d\times n_t} \triangleq \{\mathbf{x}_{n_s+1},\cdots,\mathbf{x}_{n_s+n_t}\}$, $n = n_s+n_t$, where d is dimensionality of sample vector, c is the number of emotional states and $n_{s/t}$ is the number of samples in source/target domain. Here, $\mathbf{y}^i|_{i=1}^{n_s} \in \mathbb{R}^{1\times c}$ is the label vector of $\mathbf{x}_i$ in one-hot encoding. That is, the j-th element of $\mathbf{y}^i$ is one means that the j-th sample belongs to the j-th emotional state.

The general framework of JAGP is provided in Fig. 1, which explicitly shows us that it consists of three components, *i.e.*, *feature adaptation*, *label propagation* and *graph learning*. The interactions among them are summarized as below. 1) The graph is adaptively updated based on the gradually aligned source and target EEG data, which can better guide the label propagation process (*i.e.*, emotional state estimation). 2) The graph-based semi-supervised label propagation on one hand can directly provide us with the estimated emotional states of target data; on the other hand, it can facilitate the conditional distribution modeling of target data. 3) With the help of label propagation, the source and target data can be better aligned (*i.e.*, the shared subspace can be better estimated) which in turn promotes the graph learning and label propagation. Below we describe these three components in detail to finally formulate the objective function of JAGP.

Suppose that there exists a shared subspace where the distribution discrepancies between source and target data could be minimized. Accordingly, this shared subspace is induced by a projection matrix $\mathbf{W} \in \mathbb{R}^{d\times p}$, where p is the dimensionality of the shared subspace and usually $p \ll d$. The MMD criterion [40] is adopted as the divergence measure between the projected source and target data in the shared subspace. Then, minimizing the marginal distribution discrepancy between two domains can be achieved by

$$\begin{aligned}\mathcal{M}(\mathbf{W}) &= \left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{W}^T\mathbf{x}_{s,i} - \frac{1}{n_t}\sum_{j=1}^{n_t}\mathbf{W}^T\mathbf{x}_{t,j} \right\|_2^2 \\ &= \left\| \frac{\mathbf{W}^T\mathbf{X}_s\mathbf{1}_{n_s}}{n_s} - \frac{\mathbf{W}^T\mathbf{X}_t\mathbf{1}_{n_t}}{n_t} \right\|_2^2,\end{aligned} \tag{1}$$

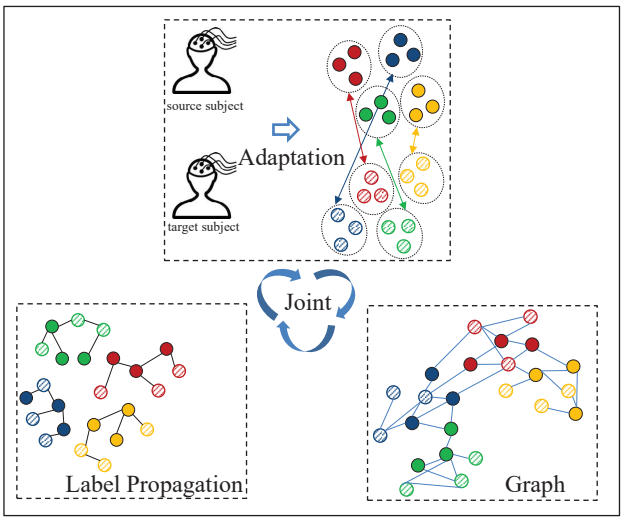


Fig. 1. The general framework of JAGP model.

where $\mathbf{x}_{s/t,i/j}$ is the i/j-th sample of $\mathbf{X}_{s/t}$ and $\mathbf{1}_{n_{s/t}}$ is an all one column vector with the size $n_{s/t}$.

However, reducing the marginal distribution divergence does not guarantee that the conditional distribution divergence between the two domains can be also reduced. Therefore, it is necessary to explicitly take the conditional distribution gap into consideration. Because the estimation of conditional distribution relies on the label information of samples but the target domain samples are unlabeled, we use the probabilistic class-wise adaptation formula [39] to perform prediction on target data. Assuming that $\mathbf{F}_t \in \mathbb{R}^{n_t \times c}$ is the estimated label indicator matrix of target samples, we construct the conditional distribution discrepancy between source and target data via MMD criterion as

$$\begin{aligned}\mathcal{C}\left(\mathbf{W},\mathbf{F}_t\right) &= \sum_{k=1}^{c}\left\|\frac{1}{n_s^k}\sum_{i=1}^{n_s^k}\mathbf{W}^T\mathbf{x}_{s,i}^k - \frac{1}{n_t^k}\sum_{j=1}^{n_t} f_t^{(j,k)}\mathbf{W}^T\mathbf{x}_{t,j}\right\|_2^2 \\ &= \left\|\mathbf{W}^T\mathbf{X}_s\mathbf{Y}_s\mathbf{N}_s - \mathbf{W}^T\mathbf{X}_t\mathbf{F}_t\mathbf{N}_t\right\|_2^2,\end{aligned} \tag{2}$$

where $\mathbf{f}_t^j \in \mathbb{R}^{1\times c}$ is the probabilistic label vector of the j-th target EEG sample. Specifically, element $f_t^{(j,k)}$ which satisfies $f_t^{(j,k)} \geq 0$ and $\sum_{k=1}^{c} f_t^{(j,k)} = 1$, represents the probability of the j-th unlabeled target EEG sample belonging to the k-th emotional state. For example, if $\mathbf{f}_t^j$=[0.07,0.12,0.79,0.02], the probabilities of the j-th sample belonging to the four emotional states are 0.07, 0.12, 0.79 and 0.02, respectively. Accordingly, this sample should be marked with the third emotional state. $n_{s/t}^k$ denotes the number of source/target domain samples belonging to k-th emotional state. Since the exact sample size of each emotional state in target domain is unavailable, we approximately calculate n_t^k by $n_t^k = \sum_{j=1}^{n_t} f_t^{(j,k)}$. $\mathbf{N}_{s/t} \in \mathbb{R}^{c\times c}$ is a diagonal matrix and its c-th diagonal element is $\frac{1}{n_{s/t}^k}$.

By combining equations (1) and (2) together, we obtain the objective function of feature adaptation as

$$\begin{aligned}\mathcal{D}\left(\mathbf{W},\mathbf{F}_t\right) &= \left\|\mathbf{W}^T\mathbf{X}_s\bar{\mathbf{Y}}_s\bar{\mathbf{N}}_s - \mathbf{W}^T\mathbf{X}_t\bar{\mathbf{F}}_t\bar{\mathbf{N}}_t\right\|_2^2 \\ s.t.\quad & \mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W} = \mathbf{I},\end{aligned} \tag{3}$$

where $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t] \in \mathbb{R}^{d\times n}$, $\bar{\mathbf{Y}}_s = [\mathbf{1}_{n_s}, \mathbf{Y}_s] \in \mathbb{R}^{n_s\times(c+1)}$, $\bar{\mathbf{F}}_t = [\mathbf{1}_{n_t}, \mathbf{F}_t] \in \mathbb{R}^{n_t\times(c+1)}$, $\bar{\mathbf{N}}_{s/t} = diag(\frac{1}{n_{s/t}}, \mathbf{N}_{s/t}) \in \mathbb{R}^{(c+1)\times(c+1)}$ and $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{n\times n}$ is the centering matrix. As discussed in [30], the constraint in (3) would help to keep the data variance after adaptation.

Given a graph similarity matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ which characterizes the relationship of samples from both source and target domains, we can perform semi-supervised label propagation on this graph [41] to estimate the label indicator matrix $\mathbf{F}_t$ of target data by minimizing

$$\min_{\mathbf{F}_t}\sum_{i,j=1}^{n}\|\mathbf{f}^i - \mathbf{f}^j\|_2^2 s_{ij} = \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}), s.t.\ \mathbf{F}_t \geq \mathbf{0}, \mathbf{F}_t\mathbf{1} = \mathbf{1}, \tag{4}$$

where $\mathbf{F} = [\mathbf{Y}_s; \mathbf{F}_t] \in \mathbb{R}^{n\times c}$, $\mathbf{L} = \mathbf{D} - \mathbf{S} \in \mathbb{R}^{n\times n}$ is the graph Laplacian matrix. $\mathbf{D}$ is a diagonal degree matrix and its i-th diagonal element is defined as $d_{ii} = \sum_j s_{ij}$. Obviously, if s_{ij} is larger which means that samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are more similar, they share the same label with a higher probability. That is, the distance between $\mathbf{f}^i$ and $\mathbf{f}^j$ should be small and then objective function (4) could be minimized.

From objective (4), we know that the higher quality of the graph, the more accurate of the estimated target labels. The graph similarity matrix $\mathbf{S}$ can be initialized by rule-based methods such as the 'HeatKernel' or '0-1' weighting schemes [42]. Since the source and target EEG samples have obvious discrepancies in the beginning based on which the constructed $\mathbf{S}$ is not accurate enough to characterize the data relationship, we should adaptively update the graph similarity matrix instead of keeping it unchanged. To be specific, it should be adaptively learned from the shared subspace representations of source and target EEG data. Inspired by the structured graph learning theory [43], [44], we enforce $\mathbf{S}$ to satisfy the constraints of non-negativity, row normalization and constrained rank, which can be achieved by minimizing the following objective

$$\begin{aligned}&\min_{\mathbf{S}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|_2^2 s_{ij} + \gamma s_{ij}^2\right) \\ &s.t.\ \mathbf{s}^i \geq \mathbf{0}, \mathbf{s}^i\mathbf{1} = 1, \mathrm{rank}(\mathbf{L}) = n - c.\end{aligned} \tag{5}$$

In (5), the former two constraints correspond to the non-negativity and row normalization properties, which are intuitive to understand. For the third one, we expect $\mathbf{S}$ to have exact c connected components corresponding to the c emotional states; then, it becomes easy to partition EEG samples into respective classes. Mathematically, the number of connected components c in graph $\mathbf{S}$ is equal to the multiplicity of eigenvalue zero of the Laplacian matrix $\mathbf{L}$. Equivalently, we constrain the rank of $\mathbf{L}$ to be $n - c$, *i.e.*, $\mathrm{rank}(\mathbf{L}) = n - c$.

However, directly optimizing the rank constraint is intractable since it is non-convex and NP hard; therefore, we make appropriate transformations to simplify its optimization. Assuming that $\sigma_i(\mathbf{L})$ is the i-th smallest singular value of $\mathbf{L}$, it is non-negative since $\mathbf{L}$ is positive semi-definite. If we incorporate $\sum_{i=1}^{c}\sigma_i(\mathbf{L})$ into (5) as a regularizer and assign it a large enough regularization parameter, the constraint in

(5) could be approximately satisfied. According to Ky Fan's Theorem [45], we have

$$\sum_{i=1}^{c}\sigma_i\left(\mathbf{L}\right)=\min_{\mathbf{F}\in\mathbb{R}^{n\times c},\mathbf{F}^T\mathbf{F}=\mathbf{I}}\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}). \tag{6}$$

From the above analysis, by relaxing the orthogonal constraint on $\mathbf{F}$ to row normalization, we obtain the objective function of JAGP by unifying (3), (4) and (5) into a single one as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{F}_t,\mathbf{S}}&||\mathbf{W}^T\mathbf{X}_s\bar{\mathbf{Y}}_s\bar{\mathbf{N}}_s-\mathbf{W}^T\mathbf{X}_t\bar{\mathbf{F}}_t\bar{\mathbf{N}}_t||_2^2+\lambda\,\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F})\\&+\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha||\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{x}_j||_2^2s_{ij}+\gamma s_{ij}^2\right)\\ s.t.\ &\mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W}=\mathbf{I},\mathbf{F}_t\geq 0,\mathbf{F}\mathbf{1}=\mathbf{1},\mathbf{S}\geq 0,\mathbf{S}\mathbf{1}=\mathbf{1},\end{aligned} \tag{7}$$

where $\mathbf{F}=[\mathbf{Y}_s;\mathbf{F}_t]\in\mathbb{R}^{n\times c}$, λ, α, γ are three non-negative regularization parameters.

3.2 Optimization

There are three variables, $\mathbf{F}_t$, $\mathbf{W}$ and $\mathbf{S}$, in objective (7), which respectively correspond to the label matrix of target data, the projection matrix to induce the shared subspace and the graph similarity matrix. Accordingly, we propose to alternately update one variable by fixing the others and below we derive their updating rules.

■ **Update $\mathbf{F}_t$ by fixing $\mathbf{W}$ and $\mathbf{S}$.** The objective function with respect to variable $\mathbf{F}_t$ is

$$\begin{aligned}\min_{\mathbf{F}_t}&||\mathbf{W}^T\mathbf{X}_s\bar{\mathbf{Y}}_s\bar{\mathbf{N}}_s-\mathbf{W}^T\mathbf{X}_t\bar{\mathbf{F}}_t\bar{\mathbf{N}}_t||_2^2+\lambda\,\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F})\\ &s.t.\ \mathbf{F}_t\geq\mathbf{0},\mathbf{F}_t\mathbf{1}=\mathbf{1}.\end{aligned} \tag{8}$$

Since matrix $\mathbf{L}$ can be decoupled as

$$\mathbf{L}=\begin{bmatrix}\mathbf{L}_{ss},\mathbf{L}_{st}\\ \mathbf{L}_{ts},\mathbf{L}_{tt}\end{bmatrix}=\begin{bmatrix}\mathbf{D}_{ss}-\mathbf{S}_{ss},\mathbf{D}_{st}-\mathbf{S}_{st}\\ \mathbf{D}_{ts}-\mathbf{S}_{ts},\mathbf{D}_{tt}-\mathbf{S}_{tt}\end{bmatrix}, \tag{9}$$

we can rewrite objective (8) as

$$\begin{aligned}\min_{\mathbf{F}_t}&\mathrm{Tr}(\mathbf{F}_t\mathbf{N}\mathbf{F}_t^T\mathbf{Z})+2\,\mathrm{Tr}(\mathbf{F}_t\mathbf{M})+\lambda\,\mathrm{Tr}(\mathbf{F}_t^T\mathbf{L}_{tt}\mathbf{F}_t)\\ &s.t.\ \mathbf{F}_t\geq\mathbf{0},\mathbf{F}_t\mathbf{1}=\mathbf{1},\end{aligned} \tag{10}$$

where $\mathbf{N}=\mathbf{N}_t\mathbf{N}_t^T\in\mathbb{R}^{c\times c}$, $\mathbf{Z}=\mathbf{X}_t^T\mathbf{W}\mathbf{W}^T\mathbf{X}_t\in\mathbb{R}^{n_t\times n_t}$ and $\mathbf{M}=\lambda\mathbf{Y}_s^T\mathbf{L}_{st}-\mathbf{N}_t\mathbf{N}_s^T\mathbf{Y}_s^T\mathbf{X}_s^T\mathbf{W}\mathbf{W}^T\mathbf{X}_t\in\mathbb{R}^{c\times n_t}$.

We denote the i-th row of $\mathbf{F}_t$ as $\mathbf{f}_t^i\in\mathbb{R}^{1\times c}$, and $\mathbf{f}_i\in\mathbb{R}^c$ as the transpose of $\mathbf{f}_t^i$. Then, we propose to optimize $\mathbf{F}_t$ row-wisely, and the objective function related to $\mathbf{f}_i$ is

$$\min_{\mathbf{f}_i}\ \lambda d_{ii}\mathbf{f}_i^T\mathbf{f}_i-2\mathbf{f}_i^T\mathbf{b}\quad s.t.\ \mathbf{f}_i\geq 0,\mathbf{f}_i^T\mathbf{1}=1, \tag{11}$$

where d_{ii} is the i-th diagonal element of the diagonal matrix $\mathbf{D}$, $\mathbf{b}=\lambda\sum_{j=1}^{n_t}\mathbf{f}_j s_{tt_{ij}}-\sum_{j=1}^{n_t}\mathbf{N}\mathbf{f}_j z_{ij}-2\mathbf{m}_i\in\mathbb{R}^{c\times 1}$, $\mathbf{m}_i\in\mathbb{R}^{c\times 1}$ is the i-th column of $\mathbf{M}$, z_{ij} is the (i,j)-th element of $\mathbf{Z}$ and $s_{tt_{ij}}$ denotes the (i,j)-th element of matrix $\mathbf{S}_{tt}$.

By completing the squared form of $\mathbf{f}_i$, we can reformulate (11) to the following form

$$\min_{\mathbf{f}_i}\ \|\mathbf{f}_i-\frac{1}{\lambda d_{ii}}\mathbf{b}\|_2^2\quad s.t.\ \mathbf{f}_i\geq 0,\mathbf{f}_i^T\mathbf{1}=1, \tag{12}$$

which defines an Euclidean projection problem on the simplex space [46]. Below we provide the complete derivations to its optimization.

Let $\mathbf{m}=\frac{1}{\lambda d_{ii}}\mathbf{b}\in\mathbb{R}^c$, and then we have the Lagrangian function corresponding to objective (12) as

$$\mathcal{L}(\mathbf{f}_i)=\|\mathbf{f}_i-\mathbf{m}\|_2^2-\delta(\mathbf{f}_i^T\mathbf{1}-1)-\boldsymbol{\eta}^T\mathbf{f}_i, \tag{13}$$

where δ, $\boldsymbol{\eta}$ are Lagrangian multipliers. Assume that $\mathbf{f}_i^*$ is the optimal solution, and δ^*, $\boldsymbol{\eta}^*$ are the corresponding optimal multipliers, respectively. According to Karush-Kuhn-Tucker (KKT) condition, for each $t=1,2,\cdots,c$, we have the following equations

$$f_{it}^*-m_t-\delta^*-\eta_t^*=0, \tag{14}$$
$$f_{it}^*\geqslant 0, \tag{15}$$
$$\eta_t^*\geqslant 0, \tag{16}$$
$$f_{it}^*\eta_t^*=0. \tag{17}$$

Equation (14) for $\mathbf{f}_i^*$ can be rewritten as

$$\mathbf{f}_i^*-\mathbf{m}-\delta^*\mathbf{1}-\boldsymbol{\eta}^*=\mathbf{0}. \tag{18}$$

According to the constraint $\mathbf{f}_i^T\mathbf{1}=1$, we have

$$\delta^*=\frac{1-\mathbf{1}^T\mathbf{m}-\mathbf{1}^T\boldsymbol{\eta}^*}{c}. \tag{19}$$

Substituting (19) into (18), $\mathbf{f}_i^*$ can be obtained by

$$\mathbf{f}_i^*=\mathbf{m}-\frac{\mathbf{1}\mathbf{1}^T}{c}\mathbf{m}+\frac{1}{c}\mathbf{1}-\frac{\mathbf{1}^T\boldsymbol{\eta}^*}{c}\mathbf{1}+\boldsymbol{\eta}^*. \tag{20}$$

Denote $\bar{\eta}^*=\frac{\mathbf{1}^T\boldsymbol{\eta}^*}{c}$ and $\mathbf{u}=\mathbf{m}-\frac{\mathbf{1}\mathbf{1}^T}{c}\mathbf{m}+\frac{1}{c}\mathbf{1}$. Then, we have

$$\mathbf{f}_i^*=\mathbf{u}+\boldsymbol{\eta}^*-\bar{\eta}^*\mathbf{1}, \tag{21}$$

Therefore, for each $t\in\{1,2,\cdots,c\}$, we have

$$f_{it}^*=u_t+\eta_t^*-\bar{\eta}^*. \tag{22}$$

According to (14) and (22), we have $u_t+\eta_t^*-\bar{\eta}^*=(u_t-\bar{\eta}^*)_+$, where $(a)_+=\max(a,0)$. Therefore, we have

$$f_{it}^*=(u_t-\bar{\eta}^*)_+. \tag{23}$$

Obviously, if $\bar{\eta}^*$ can be obtained, then the optimal solution $\mathbf{f}_i^*$ can be calculated by the above equation. Similarly, we can rewrite (22) as $\eta_t^*=f_{it}^*+\bar{\eta}^*-u_t$ such that $\eta_t^*=(\bar{\eta}^*-u_t)_+$. Hence, the $\bar{\eta}^*$ can be obtained by

$$\bar{\eta}^*=\frac{1}{c}\sum_{t=1}^{c}(\bar{\eta}^*-u_t)_+. \tag{24}$$

According to $\mathbf{f}_i^{*T}\mathbf{1}=1$ and (23), we can define a function as

$$f(\bar{\eta})=\sum_{t=1}^{c}(u_t-\bar{\eta})_+-1, \tag{25}$$

and then the optimal $\bar{\eta}^*$ can be obtained when $f(\bar{\eta}^*)=0$. by Newton method to find the root of (25) as

$$\bar{\eta}_{k+1}=\bar{\eta}_k-\frac{f(\bar{\eta}_k)}{f'(\bar{\eta}_k)}. \tag{26}$$

It is obvious that $f(\bar{\eta})$ is a piecewise linear and monotonically increasing function. If $u_t\geq\bar{\eta}$, $f(\bar{\eta})=\sum_{t=1}^{c}u_t-\bar{\eta}-1$ and we have $f'(\bar{\eta})=-1$. If $u_t<\bar{\eta}$, we have $f(\bar{\eta})=-1$ and its derivative $f'(\bar{\eta})=0$. Therefore, $f'(\bar{\eta})$ can be obtained by counting the number of positive values in $u_t-\bar{\eta}|_{t=1}^{c}$. As a result, we summarize the above derivation to sub-problem (12) in Algorithm 3.2.

Algorithm 1 The algorithm to solve sub-problem (12)

Input: vector $\mathbf{m} \in \mathbb{R}^c$
Output: variable $\mathbf{f}_i \in \mathbb{R}^c$.
1: Compute $\mathbf{u} = \mathbf{m} - \frac{\mathbf{1}\mathbf{1}^T}{c}\mathbf{m} + \frac{1}{c}\mathbf{1}$;
2: Obtain the root $\bar{\eta}^*$ of (25) by Newton's method;
3: Calculate $f_{it}^* = (u_t - \bar{\eta}^*)_+$ for each $t \in \{1, 2, \cdots, c\}$;

■ **Update $\mathbf{W}$ with $\mathbf{F}_t$ and $\mathbf{S}$ fixed**. The objective function associated with variable $\mathbf{W}$ is

$$\begin{aligned}\min_{\mathbf{W}} &\|\mathbf{W}^T\mathbf{X}_s\bar{\mathbf{Y}}_s\bar{\mathbf{N}}_s - \mathbf{W}^T\mathbf{X}_t\bar{\mathbf{F}}_t\bar{\mathbf{N}}_t\|_2^2 \\ &+ \alpha\sum_{i,j=1}^{n}\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2 s_{ij} \\ &s.t. \quad \mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W} = \mathbf{I}.\end{aligned} \tag{27}$$

By denoting $\mathbf{A} = \mathbf{X}_s\bar{\mathbf{Y}}_s\bar{\mathbf{N}}_s - \mathbf{X}_t\bar{\mathbf{F}}_t\bar{\mathbf{N}}_t \in \mathbb{R}^{d\times(c+1)}$, objective (27) can be simplified as

$$\begin{aligned}\min_{\mathbf{W}} &\|\mathbf{W}^T\mathbf{A}\|_2^2 + 2\alpha\,\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \\ &s.t.\ \mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W} = \mathbf{I}.\end{aligned} \tag{28}$$

Its Lagrangian function is defined as

$$\begin{aligned}\mathcal{L}(\mathbf{W}) = &\,\mathrm{Tr}(\mathbf{A}^T\mathbf{W}\mathbf{W}^T\mathbf{A}) + 2\alpha\,\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \\ &+ \mathrm{Tr}((\mathbf{I} - \mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W})\mathbf{\Phi}),\end{aligned} \tag{29}$$

where $\mathbf{\Phi}$ is an Lagrangian multiplier. Taking the derivative of $\mathcal{L}(\mathbf{W})$ with respect to $\mathbf{W}$ and setting it to zero, we have

$$(\mathbf{A}\mathbf{A}^T + 2\alpha\mathbf{X}\mathbf{L}\mathbf{X}^T)\mathbf{W} = \mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W}\mathbf{\Phi}. \tag{30}$$

Obviously, this is an eigenvalue decomposition problem and $\mathbf{W}$ is formed by the eigenvectors of $(\mathbf{X}\mathbf{H}\mathbf{X}^T)^{-1}(\mathbf{A}\mathbf{A}^T + 2\alpha\mathbf{X}\mathbf{L}\mathbf{X}^T)$ corresponding to its p smallest eigenvalues.

■ **Update $\mathbf{S}$ by fixing $\mathbf{F}_t$ and $\mathbf{W}$**. We can rewrite objective function (7) as

$$\begin{aligned}\min_{\mathbf{S}} &\lambda\,\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \sum_{i,j=1}^{n}\left(\alpha\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2 s_{ij} + \gamma s_{ij}^2\right) \\ &s.t.\ \mathbf{s}^i \geq \mathbf{0}, \mathbf{s}^i\mathbf{1} = 1,\end{aligned} \tag{31}$$

where $\mathbf{s}^i$ is the i-th row of $\mathbf{S}$. Since the first term in (31) can be decomposed into

$$\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{f}^i - \mathbf{f}^j\|_2^2 s_{ij}, \tag{32}$$

objective function (31) can be changed to

$$\begin{aligned}\min_{\mathbf{s}^i} &\sum_{i,j=1}^{n}\left(\frac{\lambda}{2}\|\mathbf{f}_i - \mathbf{f}_j\|_2^2 s_{ij} + \alpha\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2 s_{ij} + \gamma s_{ij}^2\right) \\ &s.t. \quad \mathbf{s}^i \geq \mathbf{0}, \mathbf{s}^i\mathbf{1} = 1.\end{aligned} \tag{33}$$

Denote $d_{ij}^{wx} = \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2$, $d_{ij}^f = \|\mathbf{f}^i - \mathbf{f}^j\|_2^2$ and $\mathbf{d}^i \in \mathbb{R}^{1\times n}$ as a vector whose j-th element is $d_{ij} = \alpha d_{ij}^{wx} + \frac{\lambda}{2}d_{ij}^f$. Considering that $\mathbf{S}$ is updated in row-wise manner, the objective function for $\mathbf{s}^i$ is

$$\begin{aligned}&\min_{\mathbf{s}^i\geq\mathbf{0},\mathbf{s}^i\mathbf{1}=1}\sum_{j=1}^{n} d_{ij}^w s_{ij} + \gamma s_{ij}^2 \\ \Leftrightarrow &\min_{\mathbf{s}^i\geq\mathbf{0},\mathbf{s}^i\mathbf{1}=1}\lambda\mathbf{s}^i(\mathbf{s}^i)^T + \mathbf{d}_i(\mathbf{s}^i)^T \\ \Leftrightarrow &\min_{\mathbf{s}^i\geq\mathbf{0},\mathbf{s}^i\mathbf{1}=1}\left\|\mathbf{s}^i + \frac{1}{2\lambda}\mathbf{d}^i\right\|_2^2.\end{aligned} \tag{34}$$

Obviously, this objective function shares the same form with (12) and thus the optimal solution $\mathbf{s}^{i*}$ can be obtained by similar derivations when optimizing $\mathbf{f}_i$.

As a whole, we summarize the optimization procedure of JAGP objective function in Algorithm 2.

Algorithm 2 The optimization to JAGP objective function

Input: Labeled EEG samples from source subject(s) $(\mathbf{X}_s, \mathbf{Y}_s)$, unlabeled EEG samples from target subject $\mathbf{X}_t$, regularization parameters λ, α, γ and the shared subspace dimension p;
Output: Projection matrix $\mathbf{W}$, predicted label matrix $\mathbf{F}_t$ of target data, and the graph similarity matrix $\mathbf{S}$.
1: Initialize the graph similarity matrix $\mathbf{S}$ by the optimal solution to problem (5) without the rank constraint, and $\mathbf{F}_t = \mathbf{1}./c \in \mathbb{R}^{n_t\times c}$;
2: Calculate the Laplacian matrix by $\mathbf{L} = \mathbf{D} - \mathbf{S}$;
3: Calculate the projection matrix $\mathbf{W}$ by (30);
4: **while** not converged **do**
5: Update $\mathbf{F}_t$ by optimizing (12) with Algorithm 3.2;
6: Update $\mathbf{W}$ by performing eigenvalue decomposition on $(\mathbf{X}\mathbf{H}\mathbf{X}^T)^{-1}(\mathbf{A}\mathbf{A}^T + 2\alpha\mathbf{X}\mathbf{L}\mathbf{X}^T)$;
7: Update $\mathbf{S}$ by solving (34) and then update the Laplacian matrix $\mathbf{L} = \mathbf{D} - \frac{\mathbf{S}^T+\mathbf{S}}{2}$;
8: **end while**

Below we analyze the computational complexity of JAGP based on the big $\mathcal{O}$ notation. Obviously, the complexity is mainly from the while-loop in Algorithm 2. For each $i \in \{1, 2, \cdots, n\}$, we need $\mathcal{O}(dp)$ to update $\mathbf{s}^i$. Then, the complexity of updating $\mathbf{S}$ is $\mathcal{O}(npd)$. The preparation of matrix $\mathbf{A}$ requires $\mathcal{O}(dn_sc+dc^2+dn_tc)$ complexity. The complexity of the line 6 in Algorithm 2 is $\mathcal{O}(d^2n+dn^2+d^3+d^2c)$. When updating each row of $\mathbf{F}_t$, the complexities of preparing intermediate variables $\mathbf{N}$, $\mathbf{Z}$ and $\mathbf{M}$ are respectively $\mathcal{O}(c^3)$, $\mathcal{O}(n_tdp+n_t^2d)$ and $\mathcal{O}(cn_sn_t+c^3+cn_sd+cdp+cdn_t)$. The complexity of using Algorithm 3.2 to update $\mathbf{F}_t$ is $\mathcal{O}(n_tc)$. Usually, we have $n > n_s > n_t \gg d > p \gg c$ in the transfer learning-based cross-subject emotion recognition. Therefore, the overall complexity of optimizing the JAGP objective function is $\mathcal{O}(tn^2d)$, where t is the number of iterations.

4 Experiments

In this section, we aim to answer the following three questions by experiments. 1) How much performance improvement will the triple unification mode of JAGP achieve in comparison with the state-of-the-art models? 2) What is the role of optimal graph learning in characterizing the relationship between pairs of EEG samples? 3) How can we identify

the common activation patterns across subjects in emotion recognition by investigating the feature weighting capacity of the learned projection matrix? The source codes of JAGP are available from https://github.com/SunsealU/JAGP.

4.1 Data Sets

Two benchmark EEG data sets, SEED_IV and SEED_V, were used in the experiments [4], [47]. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of Shanghai Jiao Tong University under Protocol No.2017060. Below we first provide general descriptions to SEED_IV and then point out the differences between them.

SEED_IV comprises 15 subjects and each subject participated the EEG data collection experiment at three different times, corresponding to three sessions. In each session, 24 video clips were sequentially displayed to evoke the four different types of emotional states, *i.e.*, *sad*, *fear*, *happy* and *neutral*. That is, six video clips correspond to one emotional state. The EEG data collection paradigm is shown in Fig. 2(a) from which we know that there are a five-second hint stage and a 45-second self-assessment stage respectively before and after the video displaying in each trial. During watching video clips, EEG data was recorded from subjects by the ESI NeuroScan system with a 62-channel electrode cap according to the international 10-20 system placement. The sampling rate is 1000 Hz.

Each session is sliced into four-second non-overlapping segments, each of which is regarded as one sample for model training. After EEG data was collected, it was first down-sampled to 200 Hz and then bandpass filtered to 1-75 Hz for artifact removal. In our experiments, the differential entropy feature [48], [49], [50] was extracted within each segment from five frequency bands, *i.e.*, *Delta* (1-3 Hz), *Theta* (4-8 Hz), *Alpha* (8-14 Hz), *Beta* (14-31 Hz), and *Gamma* (31-50 Hz). By concatenating the 62 values (corresponding to the 62 EEG channels) of each of the five frequency bands together, we obtained the sample vectors whose dimensionality is 310. Because the time durations of different video clips are slightly different, the three sessions generate different numbers of EEG samples, *i.e.*, 851, 832 and 822, respectively.

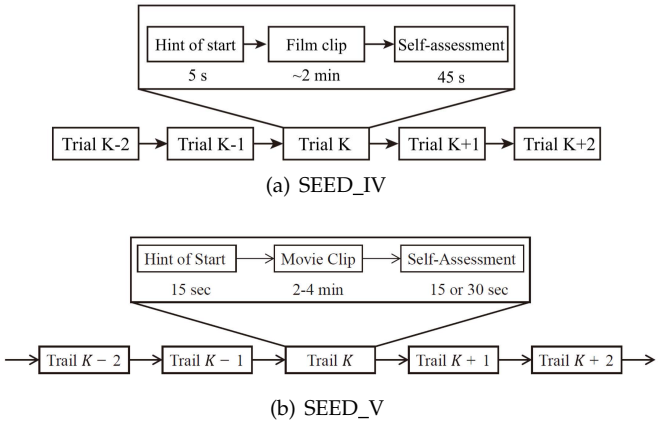


Fig. 2. The EEG data collection paradigms in SEED_IV and SEED_V.

The differences in SEED_V are listed as follows. 1) 20 subjects were recruited in the data collection experiments. The EEG data of 16 out of the 20 subjects were made public. 2) In each session, 15 video clips were displayed to elicit the five different types of emotional states, *i.e.*, *fear*, *sad*, *neutral*, *happy* and *disgust*. That is, three video clips correspond to one emotional state. The EEG data collection paradigm in SEED_V data set is provided in Fig. 2(b). 3) The number of EEG samples in the three sessions are 681, 541 and 601.

4.2 Experimental Settings

To evaluate the effectiveness of JAGP, we compare it with several state-of-the-art emotion recognition models including one semi-supervised classification model, six non-deep transfer learning models and two deep learning models. They can be roughly categorized into three groups. 1) The Gaussian Field and Harmonic Function (GFHF) [41] is a graph-based semi-supervised label propagation method which does not consider the distribution discrepancies between the source and target domains, serving as the baseline method. 2) The six non-deep transfer learning models are GFK, TCA, JDA, MIDA, GAKT and the supervised MS-STM. The former three are classical transfer learning models while the latter three are relatively new. MS-STM is a multisource transfer model by differentiating the contributions of different source subjects. GAKT tightly couples the processes of domain adaptation and target domain label estimation together, based on which JAGP is proposed by dynamically updating the graph with iteratively optimizing the model. 3) Two deep learning models are the DGCNN and LRS. Both DGCNN and JAGP use the graph data structure to model the relationship of EEG data. LRS is a typical deep transfer learning model to minimize the data discrepancies of source and target data in the latent representation similarity.

On the control experiments, we employ two transfer paradigms, *i.e.*, 'one-to-one' and 'multi-to-one'. In the former paradigm, the labeled EEG samples from subject 1 are always used as the source domain and the unlabeled EEG samples from each of the remaining subjects serve as the target domain. In the latter paradigm, when one subject serves as the target, all the remaining subjects form the source. For example, if subject 1 in SEED_IV is the target, subjects 2 to 15 form the source domain. Similar to some existing studies [6], [9], [28], we uniformly selected $1/3$ samples from these source subjects to reduce the computational burden. As a result, the sample sizes of the source domains are respectively 3976, 3892 and 3836, corresponding to the three sessions in SEED_IV. Similarly, the sample sizes of the source domains are respectively 3405, 2705, and 3006, corresponding to the three sessions in SEED_V. Since MS-STM is a multisource transfer learning model, it is only involved in the multi-to-one comparison. It is worth mentioning that except MS-STM, all the other models do not take the differences among these source subjects into consideration.

For GFHF, we adopted the k-nearest neighbor graph with 'HeatKernel' weighting scheme in which the neighborhood size k was searched from $\{5, 10, \cdots, 100\}$, and the bandwidth parameter was set as the average of all sample-pair distances. The regularization parameter λ in GFHF was searched from $\{10, 100, 1000\}$. For GFK, we tuned the dimension parameter h by grid search from $\{10, 20, \cdots, 100\}$. Linear kernel was used in both TCA and

JDA. The dimension h of the shared subspace was tuned from $\{10, 20, \cdots, 100\}$ and the parameter λ was searched from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. For GAKT, the dimension p of the shared subspace was searched from $\{10, 20, \ldots, 100\}$. In addition, there have two parameters λ, α in GAKT which were tuned from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. In MIDA, linear kernel was used. The regularization parameter μ and kernel parameter γ were tuned from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. For MS-STM, we used its supervised version and the mapping destinations were determined by the nearest prototype. Besides, the related parameters were set as suggested by [28]. In JAGP, when initializing $\mathbf{S}$, we also adopted the k-nearest neighbor graph ($k = 5$) with 'HeatKernel' weighting scheme based on the data in original space. Then, we tuned the parameters λ, α, γ from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ and the dimension p of the shared subspace was tuned with grid search from $\{10, 20, \ldots, 100\}$. Since MS-STM requires a small number of labeled target samples for model calibration, we selected the first 10 samples of each emotional state in the target domain as the labeled samples. Then, MS-STM tries to make prediction on the unlabeled samples as accurate as possible.

4.3 Recognition Results and Analysis

The cross-subject emotion recognition results of these models are respectively shown in Tables 1, 2, 3 and 4, where the best accuracy in each case is highlighted in bold. From these results, we have the following observations.

- JAGP achieved the best performance among these compared models in most cases, demonstrating the effectiveness of our triple unification of the domain-invariant feature learning, emotional label estimation and optimal graph learning. On the SEED_IV data set, JAGP obtained the average accuracy 76.75% in the one-to-one transfer paradigm, which corresponds to all the 42 test cases in Table 1. The average accuracy of the multi-to-one transfer paradigm is 78.77%, corresponding to the 45 test cases in Table 2. Similarly, on the SEED_V data set, JAGP achieved the recognition rates of 67.20% and 75.43% respectively, corresponding to the two transfer paradigms. Based on these results, we conclude that JAGP is more effective in suppressing the EEG data discrepancies than the other models, leading to better cross-subject recognition performance. Moreover, we find the recognition performance in the multi-to-one paradigm is significantly better than that in the one-to-one paradigm.
- Though both semi-supervised learning and transfer learning are useful for alleviating the small sample problem in EEG-based BCIs [51], transfer models are generally more competitive by borrowing useful information from auxiliary domain to facilitate the learning in target domain. From the obtained results, we find that JDA, GAKT, MIDA, MS-STM and JAGP are better than GFHF in most cases. Among these transfer learning models, JDA considers the alignment of both marginal and conditional distributions; however, the soft labels of the target data can be estimated by any base classifiers (*e.g.*, nearest neighbor classifier, support vector machine), which is loosely coupled with the feature adaptation. By contrast, GAKT tightly couples the two building blocks of feature adaptation and fixed graph-based label propagation. Further, the optimal graph learning is involved in JAGP, which better adapts to the gradually aligned source and target data.
- By pairwisely comparing the results respectively obtained by GAKT and JAGP, it is obvious that JAGP significantly outperforms GAKT. The main difference between them is the additional incorporation of the optimal graph learning in JAGP, which depicts the relationship among samples more accurately. However, the GAKT graph is built on the unaligned source and target data, which is not accurate enough to capture the underlying semantic meaning of EEG data. Especially when its model training comes into latter stages, such graph might mislead the emotional state estimation of target EEG data.
- By comparing with these relatively new models including the linear ones of MIDA, MS-STM and the non-linear ones of DGCNN and LRS, JAGP also achieved improved accuracies in most of the cases. The MIDA is a filter transfer model which seeks projected samples by maximizing the HSIC objective without considering the target labels. Similarly, the supervised MS-STM uses few labeled data for calibration, which does not involve the target label estimation in the transfer learning process. DGCNN is not a transfer model which focuses mainly on capturing the relationship among different EEG channels. In LRS, the marginal and conditional distributions are adapted by shallow and deep layers sequentially rather than jointly, which approximates the joint distribution in a loose way.

To investigate the statistical significance between JAGP and each of the other models, we conducted the paired t-test on their recognition results. The underlying hypothesis is that "the emotion recognition performance obtained by JAGP is better than that obtained by the other model". Each test was performed on the two accuracy sequences of JAGP and the other given model. For example, Table 1 corresponds to the 'SEED_IV: one-to-one' transfer paradigm, in which each model has the 42 values to form an accuracy sequence. Table 5 presents the statistical test results from which we find that all the elements are ✓s, meaning that the hypothesis is correct (true) with probability 0.95. For example, the notation ✓ corresponding to the case of 'SEED_IV1: JAGP>GFHF' means that the performance of JAGP is significantly better than that of GFHF in the 'SEED_IV: one-to-one' transfer paradigm. Therefore, we conclude that the declaration of 'JAGP achieves better recognition accuracies than the other models' is well supported by the experimental results.

To provide more insights into the recognition results than the accuracies only, we re-organize them in the form of confusion matrices. Due to the page limit, we only show the confusion matrices of JAGP in Fig. 3. From this figure, we can easily identify the recognition rate of JAGP on each emotional state. Moreover, the rates of samples belonging to

TABLE 1
The emotion recognition results using the 'one-to-one' transfer paradigm on SEED_IV.

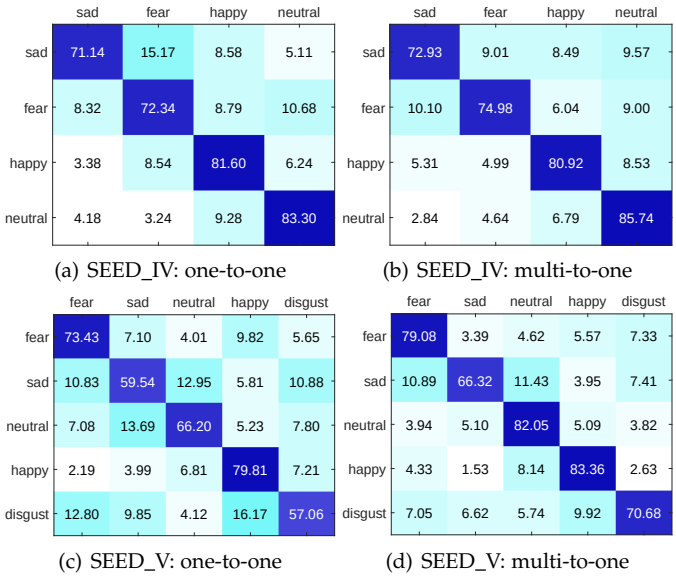
session1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	Avg.
GFHF	40.07	59.11	56.52	39.84	44.89	53.58	50.65	61.57	45.83	55.82	52.53	47.36	55.64	36.90	50.02
GFK	62.75	56.05	44.30	25.38	33.61	48.88	50.76	53.58	51.70	49.94	31.37	43.83	40.42	31.61	44.58
TCA	47.59	48.06	66.39	33.84	38.78	59.81	43.36	46.06	47.24	45.01	33.27	52.53	50.53	37.13	46.40
JDA	57.81	64.75	68.27	48.53	51.59	**70.15**	65.45	64.86	65.69	51.94	54.29	62.98	55.58	69.10	60.79
GAKT	73.09	66.75	65.92	48.18	53.11	58.87	62.51	65.22	54.29	62.16	59.34	64.28	65.45	57.34	61.18
MIDA	67.69	58.40	44.77	46.53	47.83	54.99	66.39	53.35	63.81	59.34	59.11	50.65	43.95	46.65	55.14
DGCNN	54.64	57.46	58.99	49.12	40.42	48.18	51.12	62.98	42.66	51.00	55.93	52.29	53.23	53.82	52.27
LRS	49.12	39.25	42.89	32.67	21.39	42.66	47.59	43.24	46.77	42.42	63.34	33.49	40.89	33.73	41.39
JAGP	**74.27**	**72.97**	**70.74**	**81.90**	**69.33**	66.98	**67.92**	**84.49**	**76.97**	**69.33**	**74.74**	**70.15**	**61.34**	**82.02**	**73.08**
session2	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	Avg.
GFHF	50.24	54.21	47.36	39.18	50.00	65.02	57.69	36.18	42.31	40.63	39.42	61.66	45.67	45.91	48.25
GFK	66.95	69.83	29.81	38.82	31.85	58.65	65.14	20.79	53.00	43.87	35.22	45.19	32.21	44.23	45.40
TCA	58.05	62.50	34.98	44.47	39.90	42.31	51.20	50.96	40.63	46.03	42.19	60.46	38.58	48.92	47.23
JDA	90.75	69.59	60.49	58.89	61.78	64.54	78.49	59.13	41.11	63.58	56.49	62.98	46.51	77.76	63.72
GAKT	68.03	61.54	**79.57**	63.22	57.09	68.87	68.63	54.33	**82.33**	**72.00**	52.88	64.90	52.28	88.82	66.75
MIDA	66.83	69.23	63.82	**71.03**	41.47	69.59	66.35	60.46	62.14	51.58	41.11	53.37	49.04	55.53	58.68
DGCNN	65.87	68.99	59.38	56.13	52.28	64.54	68.39	54.81	60.34	53.00	47.72	49.16	**61.66**	60.46	58.77
LRS	78.13	80.41	31.85	55.05	36.18	52.04	49.76	37.02	59.38	42.91	27.76	58.41	52.28	57.57	51.34
JAGP	**91.35**	**83.77**	69.11	68.03	**76.32**	**87.26**	**81.49**	**75.00**	78.37	71.51	**76.44**	**70.79**	61.18	**95.67**	**77.59**
session3	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	Avg.
GFHF	33.94	47.57	57.06	39.17	50.61	50.73	51.99	67.40	50.49	59.12	50.24	48.30	42.34	55.23	50.30
GFK	44.16	35.28	44.65	47.93	44.77	47.57	52.92	40.15	60.58	70.92	30.05	35.77	33.82	31.14	44.27
TCA	45.13	36.13	41.36	55.23	45.38	47.20	56.45	60.46	50.85	63.14	45.62	53.89	67.15	37.96	50.43
JDA	54.62	64.11	57.66	63.75	57.66	66.99	62.41	75.18	51.09	57.06	45.50	55.72	56.45	70.32	59.89
GAKT	60.10	65.57	69.34	67.64	62.65	79.93	59.85	50.24	69.34	**81.75**	57.54	61.44	**77.25**	85.28	67.71
MIDA	87.96	76.76	43.92	74.33	57.42	47.49	76.64	50.97	41.73	54.14	56.69	46.59	57.06	52.19	58.85
DGCNN	64.60	49.51	56.08	46.35	72.14	59.49	69.10	50.61	50.24	61.92	59.37	50.00	53.41	54.01	56.92
LRS	55.96	49.27	43.19	39.05	41.85	18.86	58.76	37.35	45.13	58.76	54.62	42.70	27.01	23.60	42.58
JAGP	**78.47**	**89.54**	**82.85**	**78.35**	**87.23**	**91.61**	**80.51**	**80.90**	**74.21**	76.03	**66.06**	**68.61**	72.38	**87.47**	**79.59**

Here, 'sub2' to 'sub15' are the indices of subjects and the results are obtained by always setting the labeled samples from subject 1 as the source.

TABLE 5
Statistical tests between JAGP and each of the other models.

	SEED_IV1	SEED_IV2	SEED_V1	SEED_V2
JAGP>GFHF	✓	✓	✓	✓
JAGP>TCA	✓	✓	✓	✓
JAGP>GFK	✓	✓	✓	✓
JAGP>JDA	✓	✓	✓	✓
JAGP>GAKT	✓	✓	✓	✓
JAGP>MIDA	✓	✓	✓	✓
JAGP>MS-STM	-	✓	-	✓
JAGP>DGCNN	✓	✓	✓	✓
JAGP>LRS	✓	✓	✓	✓

SEED_IV1 and SEED_IV2 respectively denote the 'one-to-one' and 'multi-to-one' paradigms on SEED_IV. Similarly, we have SEED_V1 and SEED_V2.



(a) SEED_IV: one-to-one (b) SEED_IV: multi-to-one

(c) SEED_V: one-to-one (d) SEED_V: multi-to-one

Fig. 3. The emotion recognition results of JAGP represented by confusion matrices.

one emotional state being misclassified into the others are also known. Taking Fig. 3(a) as an example which corresponds to the 'SEED_IV: one-to-one' paradigm, 83.3% of the *neutral* EEG samples were correctly classified by JAGP, which is the highest recognition rate among the four emotional states. Moreover, it is observed that 4.18%, 3.24%, 9.28% of the *neutral* EEG samples were wrongly recognized as the *sad*, *fear*, and *happy* states, respectively.

4.4 Insights to the Model Execution

Below we provide insights into the JAGP model execution from the two aspects.

1) The effect of feature adaptation. In JAGP, we proposed a probabilistic class-wise adaptation to reduce the discrepancies between source and target EEG data. We selected two example cases respectively corresponding to the 'one-to-one' and the 'multi-to-one' transfer paradigms on SEED_IV to intuitively show the feature adaptation ability of JAGP. The t-SNE method [52] was used to visualize the 2D representations of source and target data in both the original

TABLE 2
The emotion recognition results using the 'multi-to-one' transfer paradigm on SEED_IV.

session1	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	Avg.
GFHF	42.66	51.47	56.05	64.98	40.42	45.12	51.00	62.63	50.76	45.83	52.53	40.19	40.66	46.30	48.77	49.29
GFK	38.31	55.46	47.83	39.60	41.25	43.60	30.90	36.43	64.51	48.77	37.84	45.01	38.07	57.70	35.37	44.04
TCA	34.20	56.64	35.49	49.00	46.89	31.49	29.73	46.30	40.66	25.15	46.77	21.74	28.20	38.66	24.21	37.01
JDA	50.76	54.41	56.05	57.23	51.12	51.94	58.40	55.93	51.94	45.01	55.58	55.82	69.45	48.65	56.29	54.57
GAKT	55.93	65.33	63.45	51.70	48.41	51.59	54.41	62.04	71.33	56.64	65.10	55.46	60.87	55.11	59.93	58.49
MIDA	55.70	56.17	66.75	54.99	**76.50**	41.60	71.92	54.52	62.16	68.16	60.28	54.52	66.63	55.23	58.17	60.22
MS-STM	68.09	69.79	66.38	63.69	46.81	46.38	69.50	48.51	66.38	72.10	58.16	58.87	76.88	58.44	51.21	61.41
DGCNN	66.98	69.57	70.39	66.04	73.68	57.58	**82.61**	64.51	71.68	**82.49**	72.39	74.27	53.94	56.99	67.80	68.73
LRS	**74.38**	49.71	64.04	54.52	51.59	39.84	74.27	58.75	68.51	82.37	72.15	59.81	63.69	43.24	79.79	62.44
JAGP	70.04	**84.02**	**82.37**	**72.86**	76.38	**61.10**	75.21	**74.03**	**85.43**	81.55	**84.84**	**76.85**	**79.20**	**81.20**	**93.07**	**78.54**
session2	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	Avg.
GFHF	52.64	51.32	45.31	36.30	48.80	48.68	64.66	62.38	50.12	45.91	63.58	39.18	45.67	67.79	90.02	54.16
GFK	42.19	44.35	42.55	53.49	36.78	40.14	48.80	54.81	37.86	43.51	40.87	34.86	38.82	38.46	41.35	42.59
TCA	47.72	16.71	31.37	47.12	31.49	41.83	23.68	37.62	15.14	32.21	50.72	27.88	28.73	25.24	25.24	32.18
JDA	60.22	41.35	52.28	54.93	45.67	42.91	44.95	59.25	47.36	50.00	48.80	57.45	52.76	51.68	59.01	51.24
GAKT	71.39	78.73	64.18	51.56	65.63	66.83	63.82	63.46	63.34	65.75	53.00	49.88	52.40	55.41	89.90	63.69
MIDA	68.03	**89.30**	83.65	74.04	69.23	66.59	76.92	69.71	53.73	**71.63**	33.89	47.72	52.76	**74.52**	82.69	67.63
MS-STM	68.77	72.87	80.21	68.48	78.15	61.14	70.67	80.21	56.60	61.14	61.29	48.83	57.92	67.74	49.41	65.56
DGCNN	**83.17**	61.78	81.25	69.35	65.02	56.97	87.14	71.03	68.03	65.63	65.14	58.53	66.23	75.48	70.31	69.67
LRS	72.60	86.06	**89.06**	56.37	79.21	**77.64**	91.95	57.21	62.86	67.07	50.00	50.72	**72.60**	52.88	87.50	70.25
JAGP	78.37	86.66	84.62	**75.96**	**82.69**	72.12	**93.51**	**75.00**	**78.00**	67.67	**71.03**	**73.56**	66.11	67.79	**97.36**	**78.03**
session3	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	Avg.
GFHF	44.04	48.42	47.20	65.57	54.01	66.91	41.85	47.45	49.03	36.01	41.00	38.81	42.82	58.39	72.14	50.24
GFK	45.74	50.49	38.44	55.23	45.50	60.83	35.89	40.27	42.82	44.65	25.43	42.82	38.56	34.55	41.12	42.82
TCA	43.67	26.76	31.39	47.81	38.93	48.05	24.21	48.54	29.44	22.75	32.73	17.40	33.70	36.25	27.13	33.92
JDA	56.33	50.00	56.33	58.76	61.19	45.62	52.19	45.50	62.29	58.15	53.16	51.70	45.26	68.98	56.33	54.79
GAKT	54.74	73.48	53.89	71.17	58.27	65.21	73.84	61.92	57.42	71.90	57.30	51.32	64.23	67.03	80.66	64.16
MIDA	73.11	75.06	74.45	74.57	**83.70**	46.35	78.83	68.13	37.96	74.21	43.19	61.44	**70.80**	82.98	**84.55**	68.62
MS-STM	56.15	67.37	61.61	75.87	78.91	52.20	74.66	57.36	50.83	59.18	**70.86**	47.19	56.16	69.95	77.69	63.73
DGCNN	77.49	81.14	50.61	78.71	75.30	**82.97**	86.62	67.52	55.60	72.63	57.79	64.23	62.41	79.81	72.38	71.01
LRS	44.04	**85.77**	76.76	64.36	67.15	33.82	68.00	56.45	29.32	61.44	51.95	59.85	66.91	74.21	62.90	60.20
JAGP	**77.62**	79.93	**85.64**	**93.55**	76.03	76.03	**95.01**	**86.13**	**73.60**	**82.00**	61.31	**74.45**	70.68	**85.77**	78.35	**79.74**

Here, for each target subject, we use the labeled samples from all the other subjects as the source.

space and the learned shared subspace in Fig. 4. It is obvious that in the original feature space, the data distributions of source and target data are significantly different. While in the learned shared subspace, the data distribution discrepancies between source and target domains are dramatically reduced, demonstrating the promising feature adaptation ability of JAGP.

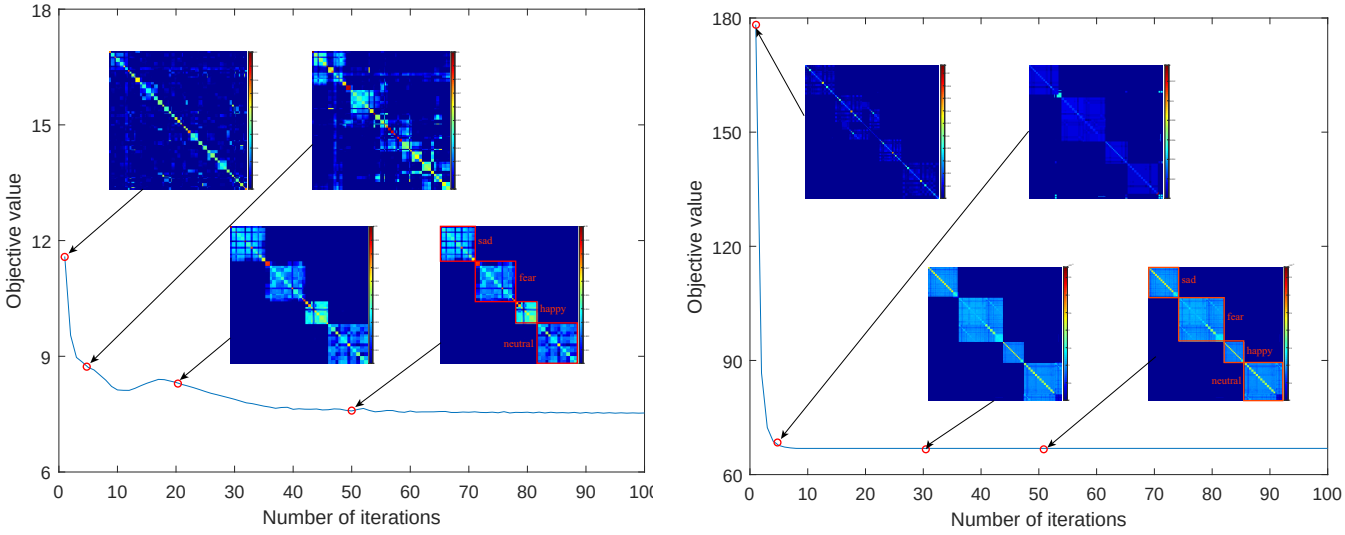
2) The optimal graph-based emotion metric learning. Unlike the GAKT, we creatively propose to learn an optimal graph from the the shared subspace representations of source and target EEG data. That is, in each iteration of model optimization, the graph similarity matrix is dynamically updated. Based on the structured graph learning theory [43], [44], in the ideal case, the number of clusters/classes should be equal to the number of connected components in graph. Taking SEED_IV for example, the number of emotional states is 4. Therefore, it is expected that the learned optimal graph should have exact four diagonal blocks, corresponding to the four emotional states.

Two example cases in SEED_IV are used to illustrate the optimal graph-based emotion metric learning. They are the 'session 2: sub3' and the 'session 2: sub15' which respectively correspond to the one-to-one and the multi-to-one paradigms. As shown in Fig. 5(a), it is observed that initially there are 48 diagonal blocks in the graph similarity matrix because each domain has exact 24 trials in SEED_IV. That is, the cross-trial differences including both cross-session and cross-subject differences are large that the graph built from unaligned data cannot well depict the semantic information (*i.e.*, emotional states) of data. As the number of iterations increases, the four diagonal blocks gradually become apparent. It means that the necessary intra-block connections are enlarged while the unnecessary inter-block connections are reduced. Since each diagonal block is a connected component corresponding to one emotional state, we consider the dynamically optimized graph as essentially learning effective emotion metric to some extent, which establishes necessary connections for EEG samples belonging to the same emotional state. Similar trends can be found from other one-to-one transfer cases.

In Fig. 5(b), we can clearly see the evolution of the four diagonal blocks as the model progresses. In this case, 810 out of all the 832 target domain samples are correctly classified. From the learned graph similarity matrix, samples belonging to the same emotional category, whether they are from the source domain or the target domain, are connected to each other. By contrast, samples from different emotion categories have no connections. This is an ideal case and the recognition rate is 97.36%. It is worth mentioning that there are multiple subjects in the source domain and we did not perform pre-selection on them. Then, it is normal that there exist large divergences between some source subjects and the target subject. Therefore, in some cases, the learned graph similarity matrices cannot guarantee to

TABLE 3
The emotion recognition results using the 'one-to-one' transfer paradigm on SEED_V.

session1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	sub16	Avg.
GFHF	41.26	45.96	54.33	20.70	45.81	25.84	55.65	39.06	20.85	47.72	34.95	50.37	37.30	54.77	52.42	41.80
GFK	36.86	37.89	29.52	31.72	31.28	28.93	47.43	43.76	19.24	44.28	48.46	36.42	44.49	43.47	39.79	37.57
TCA	42.44	36.71	47.23	23.20	39.06	29.22	33.33	49.34	23.94	37.30	38.62	58.59	39.21	42.00	43.17	38.89
JDA	49.78	40.23	49.78	22.76	45.37	49.63	36.12	54.92	31.28	44.20	44.49	44.20	49.19	43.17	50.66	43.72
GAKT	61.67	68.43	**74.89**	49.34	69.16	48.75	54.04	**69.02**	49.34	**63.29**	**63.58**	65.20	63.44	61.53	71.07	62.18
MIDA	53.01	64.32	48.60	31.86	41.26	35.98	44.93	52.86	35.39	49.05	49.78	64.02	53.60	46.26	71.66	49.51
DGCNN	44.64	54.48	51.69	37.44	48.60	37.89	43.61	55.51	36.71	43.61	46.70	55.51	50.51	50.07	60.79	47.85
LRS	34.51	40.53	59.32	34.65	32.31	23.94	39.21	28.05	28.78	31.57	45.52	71.66	37.30	35.54	**74.30**	41.15
JAGP	**61.82**	**71.37**	73.72	**54.48**	**75.48**	**54.63**	**65.20**	63.58	**49.63**	59.77	57.42	**72.54**	**65.35**	**71.66**	67.84	**64.30**
session2	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	sub16	Avg.
GFHF	36.41	39.74	56.01	49.17	56.01	52.50	55.64	49.35	34.20	40.67	49.91	58.04	37.52	34.01	55.08	46.95
GFK	33.46	42.33	36.23	40.11	34.38	50.65	51.20	34.75	31.42	22.55	32.72	41.96	23.66	45.66	54.53	38.37
TCA	31.98	32.72	40.30	47.32	46.03	61.74	50.28	27.54	31.05	30.50	34.01	52.50	17.93	37.15	50.83	39.46
JDA	35.12	38.45	51.02	45.84	31.05	61.18	64.88	39.56	45.29	51.76	44.55	59.15	17.93	22.37	57.12	44.35
GAKT	**71.35**	**64.70**	**59.33**	61.55	45.84	63.77	57.67	72.64	**57.30**	57.12	**65.25**	77.26	54.90	**67.28**	62.66	62.57
MIDA	53.05	41.05	55.82	46.21	45.29	54.53	60.44	60.63	41.40	49.17	45.29	61.37	45.10	66.91	54.34	52.04
DGCNN	40.30	42.88	46.77	32.90	40.30	59.52	46.77	53.05	46.95	35.12	39.37	54.53	42.88	38.45	58.96	45.25
LRS	29.57	40.30	36.60	34.75	23.66	38.63	52.13	48.24	31.05	32.16	41.04	55.45	30.13	48.98	**74.49**	41.15
JAGP	65.62	52.68	54.53	**73.57**	**63.22**	**76.71**	**73.57**	**82.99**	53.23	**57.30**	59.15	**88.72**	**65.43**	54.71	68.95	**66.03**
session3	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	sub16	Avg.
GFHF	34.28	58.27	59.57	63.56	37.27	25.62	49.75	34.94	44.26	59.73	54.74	46.42	52.41	48.91	44.43	47.61
GFK	32.28	40.10	45.42	48.75	36.61	51.41	51.91	50.08	24.63	45.09	42.60	52.08	44.76	33.94	56.07	43.72
TCA	38.10	46.92	43.93	52.75	42.10	36.94	58.07	48.25	37.77	42.10	54.74	51.91	60.40	38.60	60.40	47.53
JDA	46.42	46.42	56.24	55.24	40.93	43.76	89.52	57.07	35.77	60.07	54.74	63.89	58.74	58.24	51.58	54.58
GAKT	68.39	**80.87**	**64.56**	68.39	49.58	58.07	70.05	73.38	44.76	**62.56**	62.73	**78.04**	56.57	54.24	79.03	64.75
MIDA	43.26	59.07	58.40	44.76	41.43	54.08	60.23	50.08	34.11	59.73	66.06	67.55	53.58	38.77	62.40	52.90
DGCNN	54.91	53.24	50.42	41.93	49.75	45.59	50.75	62.56	42.43	44.59	42.76	58.57	55.24	51.08	56.57	50.69
LRS	37.94	64.06	44.43	43.59	36.77	53.91	66.56	57.90	22.80	30.12	54.58	54.08	49.08	38.27	56.91	47.40
JAGP	**77.54**	77.54	64.06	**72.88**	**69.38**	**69.05**	**71.21**	**94.34**	**47.59**	58.07	**70.88**	68.89	**72.71**	**74.88**	**80.20**	**71.28**



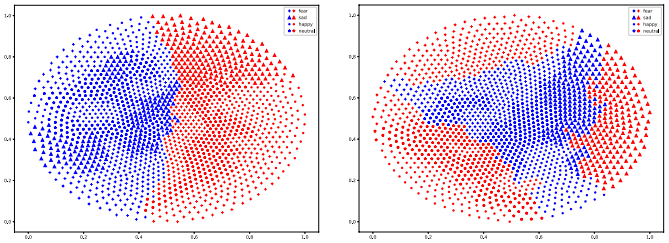
(a) session 2: sub3 in the 'one-to-one' transfer paradigm

(b) session 2: sub15 in the 'multi-to-one' transfer paradigm

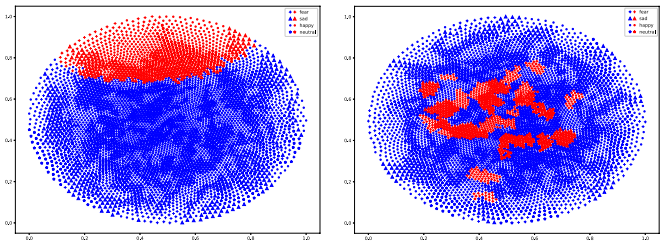
Fig. 5. Two example tasks to illustrate the optimal graph-based emotion metric learning. The four blocks respectively correspond to the four emotional states of *sad*, *fear*, *happy* and *neutral* in SEED_IV.

TABLE 4
The emotion recognition results using the 'multi-to-one' transfer paradigm on SEED_V.

session1	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	sub16	Avg.
GFHF	53.01	29.22	40.68	38.62	58.15	41.41	25.84	54.77	40.97	49.63	51.40	47.14	35.54	44.93	50.95	54.63	44.81
GFK	51.98	46.84	60.21	39.79	30.84	48.90	42.29	61.67	20.26	23.64	37.44	34.80	34.07	38.77	21.88	59.91	40.83
TCA	36.42	45.52	63.14	38.47	45.37	33.48	44.05	51.10	44.49	39.65	37.00	53.30	30.98	40.23	35.10	49.34	42.98
JDA	48.16	41.12	46.40	43.76	38.62	43.91	53.01	64.61	40.82	49.93	50.66	48.60	38.77	36.07	35.68	41.56	45.11
GAKT	72.39	58.44	71.66	67.99	50.95	**70.78**	48.02	70.78	**69.31**	44.20	63.14	60.21	64.76	59.18	62.13	**76.36**	63.14
MIDA	54.19	50.07	50.95	45.67	**63.00**	48.60	49.34	70.04	49.78	51.10	**86.78**	62.85	74.01	36.27	68.14	58.88	57.48
MS-STM	74.80	46.41	57.10	69.54	38.28	47.21	48.48	**85.49**	47.21	41.31	60.61	60.13	57.10	56.78	60.93	65.23	57.29
DGCNN	70.48	55.80	65.20	66.52	51.84	57.56	48.90	62.85	58.88	57.56	58.88	48.46	73.27	54.63	64.02	67.40	60.14
LRS	70.04	49.49	47.43	48.46	39.50	40.09	33.04	57.71	40.97	43.32	71.37	**85.90**	68.28	46.40	51.84	68.72	53.91
JAGP	**88.11**	**61.23**	**75.92**	**73.13**	57.56	67.84	**65.79**	75.77	63.29	**70.19**	80.91	69.31	**85.17**	**70.63**	**72.83**	60.79	**71.15**
session2	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	sub16	Avg.
GFHF	52.87	46.77	35.49	54.90	51.20	46.40	47.87	58.41	54.71	40.48	52.31	41.04	42.33	32.90	45.10	40.67	46.47
GFK	43.25	28.65	33.64	32.53	22.18	47.69	26.80	46.77	24.21	30.13	38.63	37.15	45.47	41.59	30.50	24.95	34.63
TCA	39.56	38.63	44.73	34.75	36.04	54.53	38.60	52.68	42.70	30.50	48.98	37.89	88.51	33.46	35.86	31.61	43.06
JDA	44.55	42.51	56.38	50.09	48.98	51.57	45.27	66.73	47.13	52.68	57.86	48.06	45.01	45.10	70.43	46.40	51.17
GAKT	**81.33**	64.14	64.14	68.21	**69.69**	47.32	64.51	76.89	75.97	52.31	67.47	68.21	81.89	54.53	50.46	60.26	65.46
MIDA	69.13	33.09	65.62	41.96	34.57	48.43	80.41	75.42	48.61	34.38	43.81	59.33	76.16	34.01	46.77	48.98	52.54
MS-STM	72.51	53.56	71.08	63.34	56.82	53.56	67.82	61.91	62.32	38.09	54.28	55.19	67.21	34.42	48.88	72.30	58.33
DGCNN	56.56	53.97	74.31	56.56	59.52	48.43	49.17	69.32	55.82	53.23	55.45	55.27	84.29	44.18	62.85	59.15	58.63
LRS	48.98	40.48	**82.26**	24.03	43.81	51.76	81.15	73.75	34.01	19.41	67.47	**90.94**	82.81	48.24	63.22	52.50	56.55
JAGP	73.57	**69.69**	80.04	**80.22**	63.03	**70.79**	**87.80**	**94.45**	**81.52**	**61.74**	**80.59**	85.03	**91.87**	**70.43**	**85.58**	**83.55**	**78.74**
session3	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	sub16	Avg.
GFHF	43.26	37.44	44.59	49.25	50.58	56.57	43.43	61.23	44.09	49.75	43.93	35.94	47.25	48.42	52.75	35.27	46.48
GFK	38.44	36.77	35.11	42.10	39.70	30.45	35.44	38.60	43.93	31.78	46.59	37.27	54.58	57.74	35.77	40.93	40.33
TCA	32.61	35.94	40.77	35.94	47.75	27.79	68.39	25.79	41.93	30.95	42.10	31.11	65.39	41.26	33.94	59.07	41.30
JDA	48.75	40.60	44.76	64.56	60.40	45.09	52.75	45.09	48.42	44.09	55.24	44.76	76.37	43.59	64.89	52.58	52.00
GAKT	65.56	56.91	71.71	77.20	59.90	**67.22**	55.07	73.21	65.72	46.59	56.07	65.56	81.03	47.42	57.57	63.73	63.15
MIDA	61.56	58.40	71.05	71.88	44.09	40.27	73.54	83.19	83.53	38.94	75.37	68.89	65.39	44.93	67.72	53.74	62.66
MS-STM	91.65	47.37	**86.21**	78.95	63.34	58.98	54.99	68.97	72.41	21.78	53.18	**72.78**	67.15	52.63	31.58	**86.57**	63.03
DGCNN	73.88	61.06	64.39	73.88	42.93	53.08	51.75	65.72	75.37	47.92	65.56	50.58	75.71	55.24	60.40	60.40	61.12
LRS	38.77	37.10	47.25	**85.36**	53.91	29.28	62.23	81.70	63.23	30.45	52.58	52.25	87.69	48.09	**83.86**	79.37	58.32
JAGP	**82.03**	**88.52**	82.03	80.53	**76.21**	47.42	**96.51**	**86.52**	**91.18**	**59.23**	**75.87**	65.89	**87.85**	**58.90**	68.22	75.37	**76.39**



(a) 2D visualization of source and target data in original feature space (left) and the learned shared subspace (right). This example corresponds to the case of session2: subject1 → subject15 in the one-to-one transfer paradigm on SEED_IV.



(b) 2D visualization of source and target data in original feature space (left) and the learned shared subspace (right). This example corresponds to the case of session1: subject9 in the multi-to-one transfer paradigm on SEED_IV.

Fig. 4. 2D visualization of the feature adaptation ability of JAGP on SEED_IV. Blue dots represent the source domain samples and red ones denote the target domain samples.

have exact four diagonal blocks. As pointed by [28], more source subjects do not necessarily lead to better recognition performance of target domain. Accordingly, domain transferability estimation is helpful to select source subjects [10], [53], which can reduce the risk of negative transfer and the computational burden.

4.5 Insights to the Learned Shared Subspace

In this subsection, we provide insights into the learned shared subspace induced by the projection matrix $\mathbf{W}$. From the perspective of transfer learning, $\mathbf{W}$ should encourage the subject-independent features and simultaneously suppress the subject-dependent features in cross-subject emotion recognition. Then, in the learned subspace, the projected representations of source and target data would be aligned as much as possible. Existing studies investigated the learned shared subspace mainly by analyzing the data distributions before and after projection [9], [10]. Here, we judge the weight of each EEG feature to be a common/shared feature across subjects and further automatically identify where these features from (*e.g.*, the EEG frequency bands and channels).

As shown in Fig. 6, we use $\mathbf{w}^i|_{i=1}^{310} \in \mathbb{R}^p$ to denote the i-th row of $\mathbf{W}$. Inspired by [54], the importance of the i-th EEG feature, θ_i, can be measured by

$$\theta_i = \frac{\|\mathbf{w}^i\|_2}{\sum_{j=1}^{d} \|\mathbf{w}^j\|_2}, \tag{35}$$

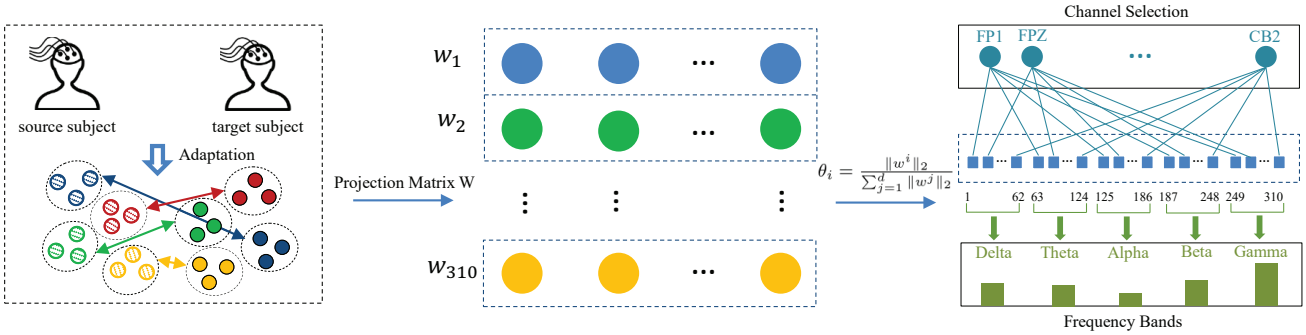


Fig. 6. Illustration to the projection matrix-based feature importance measure and critical frequency bands (channels) identification.

where $\|\cdot\|_2$ is the ℓ_2-norm defined by $\|\mathbf{w}^i\|_2 = \sqrt{\sum_{j=1}^{p} w_{ij}^2}$. Once $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_{310}]$ is calculated, we obtain the quantitative importance measures of all EEG features. The underlying mechanism is that *if the value of θ_i is larger, the i-th EEG feature is more common across subjects*. Since there exists correspondence between EEG feature dimensions and the frequency bands (channels) shown in the rightmost part of Fig. 6, we can identify the important EEG frequency bands and channels which generate these more common and important EEG features. To be specific, the importance of the $k|_{k=1}^{5}$-th frequency band can be measured by

$$\omega(k) = \theta_{(k-1)*62+1} + \theta_{(k-1)*62+2} + \cdots + \theta_{k*62}. \quad (36)$$

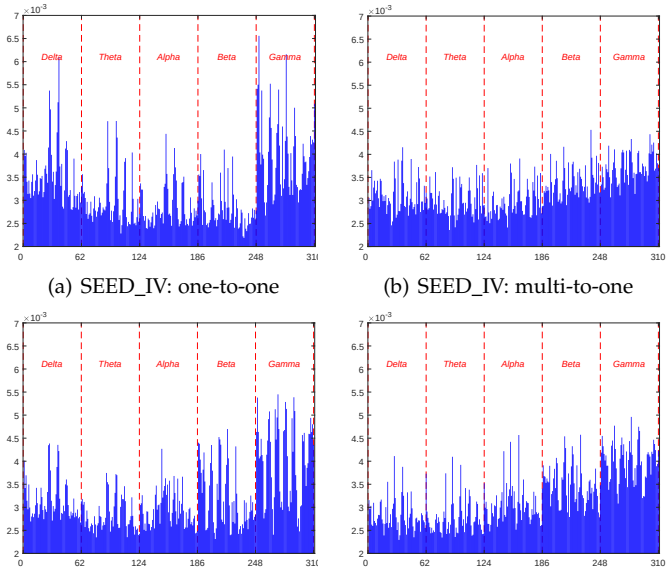
Similarly, the importance of the $q|_{q=1}^{62}$-th EEG channel can be calculated by

$$\psi(q) = \theta_q + \theta_{q+62} + \cdots + \theta_{q+248}. \quad (37)$$

It is worth mentioning that though there are 62 EEG channels and 5 frequency bands in SEED_IV and SEED_V, this method is general enough to be applied to any data set with arbitrary number of frequency bands (channels) on condition that the correspondence between EEG feature dimensions and frequency bands (channels) can be established. Then, rules (36) and (37) can be adapted to the new data set. To the best of our knowledge, this is the first time to analyze the learned shared subspace from the perspective of feature importance in transfer learning-based cross-subject emotion recognition.

When the projection matrix is learned, we obtain the importance of different EEG features based on equation (35). In Fig. 7, we show the quantitative importance of different EEG features where the heights of vertical lines are the specific values. Obviously, some of the features contribute more significantly while some contribute less. These significant features can be viewed as the subject-invariaent features, which should be encouraged by the learned projection matrix. Differently, these less significant features are corresponding to subject-dependent features which should be suppressed by the projection matrix.

Since EEG data is typically multi-rhythm and multi-channel, it is necessary to make further analysis on where these significant features are mainly from (*e.g.*, which frequency bands and channels). According to the rule (36), we sum over the corresponding values of θs to obtain the importance of each frequency band in Fig. 8. The quantitative importance values are more intuitively placed on the top of the bars. Obviously, the *Gamma* frequency band is the most important one which is considered as the main frequency band to generate the domain-invariant features, from the perspective of pattern recognition. Similarly, based on rule (37), we obtain the importance values of the 62 EEG channels, which are displayed in the form of brain topographical maps in Fig. 9. In EEG-based aBCI research, there exists a consensus that the channel-wise features in different brain regions might correlate differently to emotion expression. From our experimental results, we conclude that there are four main brain regions, *i.e.*, the prefrontal, left/right temporal, and (central) parietal lobes, which mainly generate the important features across subjects.



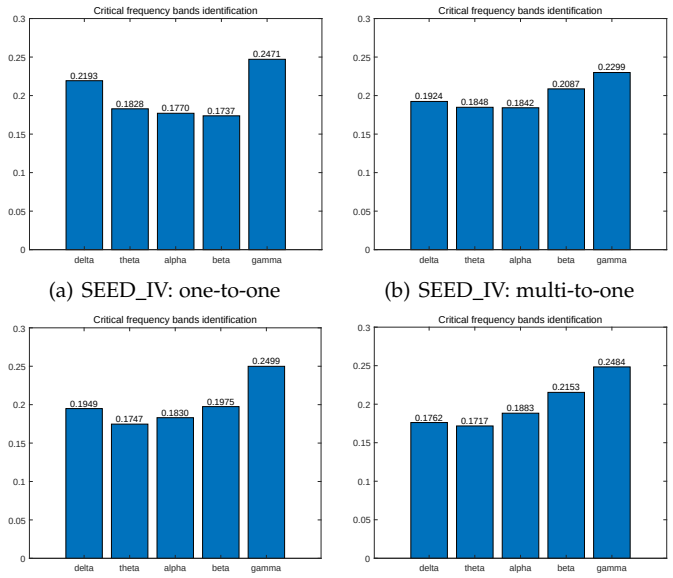
(a) SEED_IV: one-to-one (b) SEED_IV: multi-to-one

(c) SEED_V: one-to-one (d) SEED_V: multi-to-one

Fig. 7. The importance of different EEG feature dimensions.

4.6 Model Convergence and Parameter Sensitivity

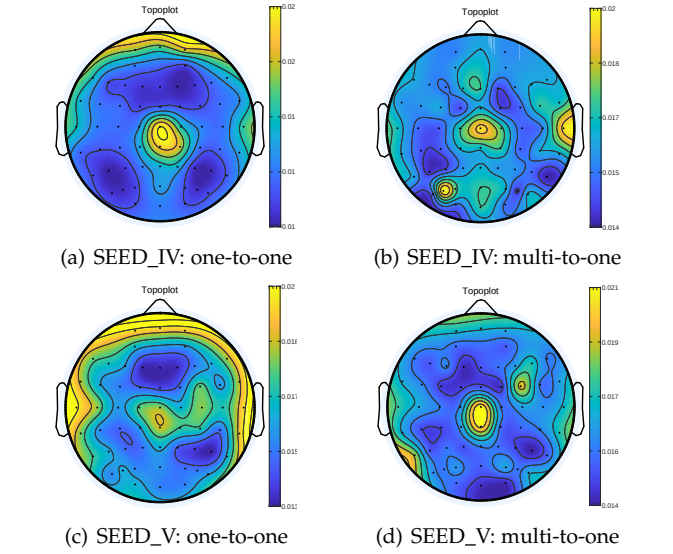
1) The convergence of JAGP. There are three blocks in the loop of Algorithm 2, corresponding to the updating of the three variables in JAGP. To experimentally test the convergence of our proposed iterative optimization algorithm, the bi-coordinate system was used to simultaneously show the decreasing of JAGP objective values and the increasing of recognition accuracies as the number of iterations increases.



(a) SEED_IV: one-to-one (b) SEED_IV: multi-to-one

(c) SEED_V: one-to-one (d) SEED_V: multi-to-one

Fig. 8. The importance of different EEG frequency bands.



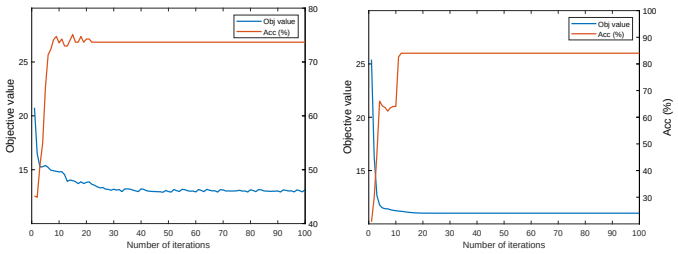
(a) SEED_IV: one-to-one (b) SEED_IV: multi-to-one

(c) SEED_V: one-to-one (d) SEED_V: multi-to-one

Fig. 9. The importance of EEG channels in brain topographical map.

In Fig. 10, two examples respectively corresponding to the one-to-one paradigm and the multi-to-one paradigm on SEED_IV are provided. It is observed that with increasing number of iterations, the objective values generally keep decreasing while the recognition accuracy keeps improving and finally achieves stability. Further, we find that JAGP converges in a few iterations, usually less than 40.

According to [46], [55], the graph corresponding to target domain samples can be reconstructed from the estimated soft labels by $\mathbf{F}_t\mathbf{F}_t^T$. In Fig. 11, we use two examples to intuitively show the evolution of JAGP in estimating $\mathbf{F}_t$. Obviously, as the number of iterations increases, the diagonal blocks in reconstructed graphs become clearer and the prediction accuracies are improved gradually. Such result reflects the effectiveness of the joint optimization mode in JAGP; specifically, the three modules of EEG feature adaptation, optimal graph learning and graph-based emotional state estimation are tightly coupled in a unified framework. Similar results can be found in other cases.



(a) session 1: sub10 (b) session 2: sub2

Fig. 10. Two examples to show the decreasing of JAGP objective values as well as the increasing of recognition accuracies on SEED_IV.

2) Parameter sensitivity analysis. We know that transfer learning aims to find a shared subspace where the divergence between source and target data is minimized. Mathematically, such shared subspace is induced by the projection matrix $\mathbf{W} \in \mathbb{R}^{d\times p}$, where d=310 is the dimensionality of the feature vector and p is the shared subspace dimensionality. To investigate how many dimensions of the shared subspace are sufficient to capture the common information of source and target EEG data, in Fig. 12(a) we draw the recognition accuracies of JAGP in terms of different ps on some example cases in SEED_IV. This on one hand reflects that there exist redundancies in the 310-dimensional representation of EEG samples and there might be more important EEG frequency bands and channels for cross-subject emotion recognition. On the other hand, the shared subspace uses a much lower dimensionality to suppress the individual differences. In Fig. 12(b), we show the average performance of JAGP in terms of subspace dimensions. As can be seen from this figure, JAGP can achieve better accuracies when the subspace dimensionality is about 30.

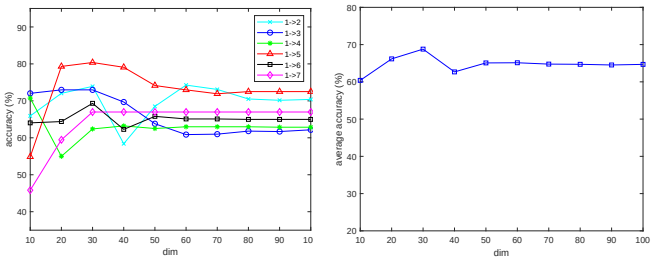
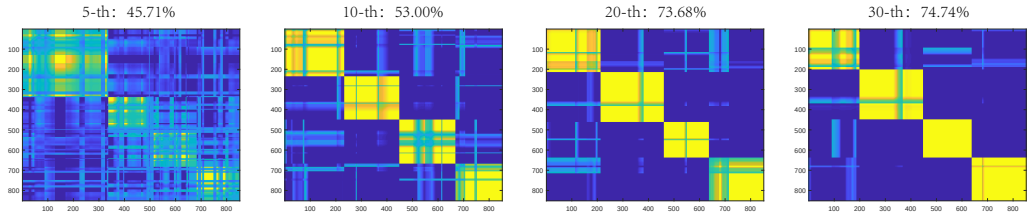
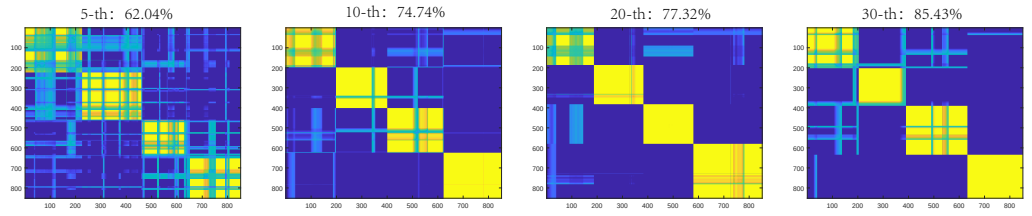


Fig. 12. The performance of JAGP in terms of subspace dimensions. Example cases (left) and the average (right).

Except the subspace dimensionality parameter, we have three regularization parameters λ, α and γ in the objective function of JAGP, among which α corresponds to the label propagation term, α and γ correspond to the optimal graph learning term. Here we investigate the performance of JAGP in terms of the variation of one parameter by fixing the two others to near optimal values. Taking SEED_IV as an example, we respectively show the average performance of JAGP versus the three parameters in the three subfigures of Fig. 13. From this figure, we find that JAGP is not sensitive to the variations of λ and α which have many candidate values to make JAGP achieve good recognition performance. By

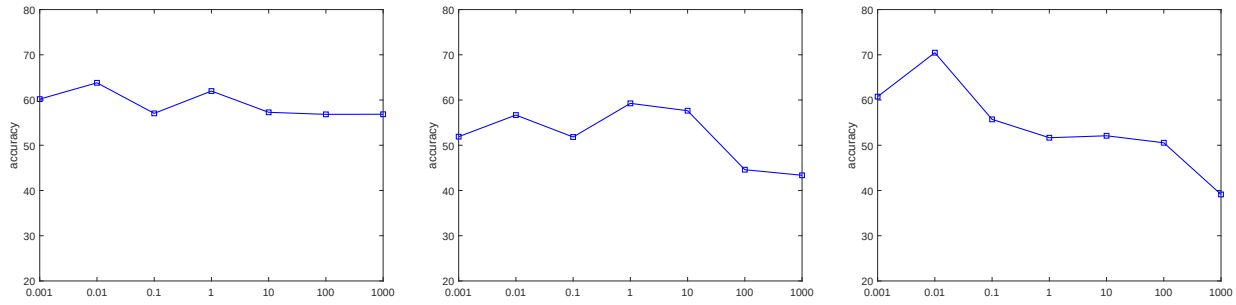


(a) session 1: sub12 in the 'one-to-one' transfer paradigm



(b) session 1: sub9 in the 'multi-to-one' transfer paradigm

Fig. 11. The reconstructed graphs by estimated target labels in terms of iterations. The four blocks are respectively corresponding to the four emotional states of *sad*, *fear*, *happy* and *neutral* in SEED_IV.



(a) Accuracy(%) vs. λs (b) Accuracy(%) vs. αs (c) Accuracy(%) vs. γs

Fig. 13. The average performance of JAGP with parameters λ, α and γ on SEED_IV.

contrast, JAGP is significantly affected by parameter γ for which 0.01 is a reasonable choice in most cases.

5 Discussions

In this section, we discuss the superiorities of the proposed JAGP model first and then the interpretability of the learned feature importance.

On the superiorities of JAGP, we summarize them from the below three aspects. 1) JAGP seamlessly integrates the three blocks of *feature adaptation*, *label propagation* and *graph learning* together to form a unified objective. It is theoretically better than the loosely-coupled transfer models which ofteh estimate the soft target labels by base classifiers, *e.g.*, k nearest neighbor classifier or support vector machine [30], [56]. In JAGP, we see that the label indicator matrix $\mathbf{F}_t$ appears in all the three blocks, the projection matrix $\mathbf{W}$ aims at finding a shared subspace for aligning source and target data, and graph similarity matrix $\mathbf{S}$ works for optimal graph learning based on which the semi-supervised emotional state estimation can be performed. Particularly, different from the real-valued setting of the label indicator matrix by existing studies [43], [44], the utilization of non-negative and row-normalization constraints on $\mathbf{F}_t$ enable us to directly obtain the emotion recognition results without postprocessing. 2) Inspired by the fact that learning performance can be greatly improved by taking the data manifold into consideration, the graph structure has been widely to characterize the relationship among data. For example, inspired by the connectivity patterns of multi-channel EEG data, in [15], graph was used to model the relationship among different EEG channels and the resultant DGCNN model showed satisfactory emotion recognition performance from EEG. In JAGP, other than constructing a fixed graph to measure the similarities of unaligned source and target EEG samples, we propose to dynamically update the graph similarity matrix according to their shared subspace representations; therefore, the relationship among data is much easier to capture and therefore more accurate graph similarity matrix is achieved to better guide the label propagation process. For EEG-based emotion recognition, an optimal graph can better establish connections among EEG samples belonging to the same emotional state and remove unnecessary connections of EEG samples belonging to different emotional states. That is, the graph adjacency matrix can be viewed as a learned emotion metric. Based on the above analysis, it is obvious that the roles of graph in DGCNN and JAGP are different. The former puts emphasis on modeling the spatial information of multi-channel EEG data while the latter measures the similarities of EEG samples. 3) On the role of the projection matrix in JAGP, it aims to find a shared

subspace where the source and target samples are expected to be aligned. This is exactly what the transfer learning models do. Besides the recognition accuracy, existing transfer learning models in EEG-based BCIs mainly compared the source and target data distributions before and after transfer by visualizing them in 2D or 3D subspace. However, few of them provided more insights into the learned shared subspace. In the present work, we investigated the feature weighting ability of the learned projection matrix by JAGP, which is associated with the identification of critical EEG frequency bands (channels) shared across subjects in emotion recognition. From the machine learning perspective, the learned projection matrix should encourage the subject-independent features and suppress the subject-dependent features in order to explore the common information shared by both source and target data. Inspired by the progresses in feature selection and ranking [54], [57], [58], we used the normalized ℓ_2-norm of each row of the projection matrix to measure the importance of each EEG feature dimension. Then, according to the established correspondence between each dimension of EEG frequency domain features and a certain frequency band (channel), we can determine where those important features are extracted from (*e.g.*, frequency bands and channels).

On the interpretability of learned feature importance, we have the following understandings. 1) In pattern recognition, feature importance lays the theoretical foundation for feature selection and feature ranking, whose underlying rationality is that features have larger importance values are more discriminative in differentiating samples of different classes. Taking the rescaled least square regression (RLSR) model as an example [57], [58], it provides us with a quantitative way to measure the feature importance by directly connecting the data matrix with label matrix. Therefore, RLSR can be used to identity the important feature dimensions of data such as the example data sets proposed in [59] [1]. 2) In JAGP, the projection matrix aims to induce the shared subspace. Inspired by RLSR, we used the normalized row ℓ_2-norm to measure the feature importance values, which depicts the common information shared across subjects. Further, the corresponding EEG frequency bands and channels can be identified. As pointed by [59], it should be careful to use the learned model weights to interpret the neural process specific to a certain mental state. Therefore, though the identification results in EEG frequency bands and channels are consistent with existing studies in emotion recognition to some extent [13], [20], [60], whether they are essentially related to the occurrence of emotion expression requires some further researches from cognitive neuroscience.

6 Conclusion

In this paper, we proposed a cross-subject emotion recognition model termed JAGP for simultaneously completing the EEG feature adaptation and graph adaptive emotional state estimation. JAGP has three advantages including unifying the three components of feature adaptation, emotional state estimation and optimal graph learning together in a tightly-coupled manner, updating the graph gradually to learn better emotion metric as the number of iterations increases, and performing systematic analysis on the learned shared subspace from the perspective of measuring the feature importance. Extensive experiments were conducted on two benchmark EEG data sets and the results demonstrated that 1) JAGP obtained improved cross-subject emotion recognition performance which benefits from the triple unification mode, 2) the gradual updating of graph effectively performs emotion metric learning by adaptively enhancing the connections of EEG samples belonging to the same emotional state, and 3) the *Gamma* band and the prefrontal, left/right temporal, (central) parietal lobes, are identified as the main frequency band and brain regions in generating common features across subjects in emotion recognition. In the future, on one hand we will extend JAGP into nonlinear version to better characterize the possible nonlinear structure of EEG data; on the other hand, we will consider domain transferability estimation for source subject selection.

[1] https://github.com/SunseaIU/Interpretation_weights

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grants 61971173 and U20B2074, National Key Research and Development Program of China under Grant 2017YFE0116800, Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant GK209907299001-008, Natural Science Foundation of Zhejiang Province under Grant LY21F030005, China Postdoctoral Science Foundation under Grant 2017M620470, CAAC Key Laboratory of Flight Techniques and Flight Safety under Grant FZ2021KF16 and Guangxi Key Laboratory of Optoelectronic Information Processing (Guilin University of Electronic Technology) under Grant GD21202.

References

[1] Y. Shu and S. Wang, "Emotion recognition through integrating EEG and peripheral signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 2017, pp. 2871–2875.

[2] Z. He, Z. Li, F. Yang, L. Wang, J. Li, C. Zhou, and J. Pan, "Advances in multimodal emotion recognition based on brain–computer interfaces," *Brain Sci.*, vol. 10, no. 10, p. 687, 2020.

[3] X.-L. Quan, Z.-G. Zeng, J.-H. Jiang, Y.-Q. Zhang, B.-L. Lu, and D.-R. Wu, "Physiological signals based affective computing: a systematic review (in Chinese)," *Acta Automatica Sinica*, vol. 47, no. 8, pp. 1769–1784, 2021.

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

[5] Y. Peng, F. Qin, W. Kong, Y. Ge, F. Nie, and A. Cichocki, "GFIL: A unified framework for the importance analysis of features, frequency bands and channels in EEG-based emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, 2021, DOI: 10.1109/TCDS.2021.3082803.

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

[7] L.-M. Zhao, X. Yan, and B.-L. Lu, "Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 863–870.

[8] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 1, pp. 85–94, 2019.

[9] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for EEG emotion recognition based on latent representation similarity," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 2, pp. 344–353, 2020.

[10] J. Cui, X. Jin, H. Hu, L. Zhu, K. Ozawa, G. Pan, and W. Kong, "Dynamic distribution alignment with dual-subspace mapping for cross-subject driver mental state detection," *IEEE Trans. Cognit. Develop. Syst.*, 2021, DOI: 10.1109/TCDS.2021.3137530.

[11] G. Zhang, M. Yu, G. Chen, Y. Han, D. Zhang, G. Zhao, and Y. Liu, "A review of EEG features for emotion recognition (in chinese)," *Scientia Sinica-Informationis*, vol. 49, no. 9, pp. 1097–1118, 2019.

[12] 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, 2014.

[13] 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, 2019.

[14] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl. Soft Comput.*, vol. 100, no. 106954, pp. 1–12, 2021.

[15] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, 2020.

[16] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "SparseDGCNN: Recognizing emotion from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, 2021, DOI: 10.1109/TAFFC.2021.3051332.

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

[18] W.-C. L. Lew, D. Wang, K. Shylouskaya, Z. Zhang, J.-H. Lim, K. K. Ang, and A.-H. Tan, "EEG-based emotion recognition using spatial-temporal representation via Bi-GRU," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 116–119.

[19] S. Gong, K. Xing, A. Cichocki, and J. Li, "Deep learning in EEG: Advance of the last ten-year critical period," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 2, pp. 348–365, 2022.

[20] 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, 2015.

[21] N. S. Suhaimi, J. Mountstephens, and J. Teo, "EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities," *Comput. Intell. Neurosci.*, vol. 2020, no. 8875426, pp. 1–19, 2020.

[22] B.-L. Lu, Y.-Q. Zhang, and W.-L. Zheng, "A survey of affective brain-computer interface (in Chinese)," *Chinese J. Intell. Sci. Techn.*, vol. 3, no. 1, pp. 36–48, 2021.

[23] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," *IEEE Trans. Artif. Intell.*, vol. 1, no. 2, pp. 151–166, 2021.

[24] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[25] A. Dong, Z. Li, and Q. Zheng, "Transferred subspace learning based on non-negative matrix factorization for EEG signal classification," *Front. Neurosci.*, vol. 15, no. 647393, pp. 1–12, 2021.

[26] X. Gu, M. Gao, Y. Jiang, X. Ning, and P. Qian, "Multi-source domain transfer discriminative dictionary learning modeling for Electroencephalogram-based emotion recognition," *IEEE Trans. Comput. Soc. Syst.*, 2022, DOI: 10.1109/TCSS.2022.3153660.

[27] X.-Y. Zhang and C.-L. Liu, "Writer adaptation with style transfer mapping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1773–1787, 2012.

[28] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3281–3293, 2020.

[29] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Net.*, vol. 22, no. 2, pp. 199–210, 2010.

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

[31] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2012, pp. 2066–2073.

[32] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 288–299, 2018.

[33] Y. Li, B. Fu, F. Li, G. Shi, and W. Zheng, "A novel transferability attention neural network model for EEG emotion recognition," *Neurocomputing*, vol. 447, pp. 92–101, 2021.

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

[35] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A novel bi-hemispheric discrepancy model for EEG emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 2, pp. 354–367, 2021.

[36] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for EEG-based brain-computer interfaces: A review of progress made since 2016," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 1, pp. 4–19, 2022.

[37] D. Wu, X. Jiang, R. Peng, W. Kong, J. Huang, and Z. Zeng, "Transfer learning for motor imagery based brain-computer interfaces: a complete pipeline," *arXiv preprint arXiv:2007.03746*, 2020.

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

[39] Z. Ding, S. Li, M. Shao, and Y. Fu, "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 37–52.

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

[41] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 912–919.

[42] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, 2010.

[43] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2014, pp. 977–986.

[44] F. Nie, X. Wang, M. Jordan, and H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1969–1976.

[45] I. Gutman, E. A. Martins, M. Robbiano, and B. San Martin, "Ky Fan theorem applied to randić energy," *Linear Algebra Appl.s*, vol. 459, pp. 23–42, 2014.

[46] Y. Peng, X. Zhu, F. Nie, W. Kong, and Y. Ge, "Fuzzy graph clustering," *Inf. Sci.*, vol. 571, pp. 38–49, 2021.

[47] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 2, pp. 715–729, 2021.

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

[49] J. Zhang, Z. Wei, J. Zou, and H. Fu, "Automatic epileptic EEG classification based on differential entropy and attention model," *Eng. Appl. Artif. Intell.*, vol. 96, no. 103975, pp. 1–10, 2020.

[50] J. Wu, *Essentials of pattern recognition: an accessible approach*. Cambridge University Press, 2020.

[51] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Inf. Sci.*, vol. 474, pp. 90–105, 2019.

[52] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[53] W. Zhang and D. Wu, "Manifold embedded knowledge transfer for brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1117–1127, 2020.

[54] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint ℓ2,1-norms minimization," *Proc. Adv. Neural Inf. Proc. Syst.*, vol. 23, pp. 1813–1821, 2010.

[55] J. Han, K. Xiong, and F. Nie, "Orthogonal and nonnegative graph reconstruction for large scale clustering," in *Proc. Int. J. Conf. Artif. Intell.*, 2017, pp. 1809–1815.

[56] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, and C. P. Chen, "Guide subspace learning for unsupervised domain adaptation," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 31, no. 9, pp. 3374–3388, 2020.

[57] X. Chen, G. Yuan, F. Nie, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression," in *Proc. Int. J. Conf. Artif. Intell.*, 2017, pp. 1525–1531.

[58] X. Chen, G. Yuan, F. Nie, and Z. Ming, "Semi-supervised feature selection via sparse rescaled linear square regression," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 165–176, 2020.
[59] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *Neuroimage*, vol. 87, pp. 96–110, 2014.
[60] Y. Peng, W. Kong, F. Qin, F. Nie, J. Fang, B.-L. Lu, and A. Cichocki, "Self-weighted semi-supervised classification for joint EEG-based emotion recognition and affective activation patterns mining," *IEEE Trans. Instrum. Meas.*, vol. 70, no. 2517111, pp. 1–11, 2021.

Yong Peng (M'18) received the BS degree in computer science from PLA Artillery Academy, Hefei, China, in 2006, the MS degree in computer science from Graduate University of Chinese Academy of Sciences, Beijing, China, in 2010, and the PhD degree in computer science from Shanghai Jiao Tong University, Shanghai, China in 2015. Currently he is a full professor with School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His main research interests are machine learning, pattern recognition and EEG-based brain-computer interfaces. He has published more than 30 SCI(SSCI)-indexed journal papers such as IEEE TNSRE, TCDS, TIM, TCAS-II, Information Sciences, Neural Networks, and Knowledge-Based Systems. He was awarded by the President Prize of Chinese Academy Sciences in 2009 and the Third Prize of Chinese Institute of Electronics in 2018.

Wenjuan Wang received the BS degree in computer science from Hefei University, Hefei, China, in 2019. She is currently pursuing the master degree in School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. Her current research interests include machine learning and EEG-based brain-computer interfaces.

Wanzeng Kong (M'13) received the BS degree in automation from Zhejiang University, Hangzhou, China, in 2003 and the PhD degree in control theory and control engineering Zhejiang University, Hangzhou, China, in 2008. He was a Visiting Research Associate with the Department of Biomedical Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA, from 2012 to 2013. He is currently a full Professor with School of Computer Science and Technology, Hangzhou Dianzi University and the Dean of Zhejiang Key Laboratory of Brain-Machine Collaborative Intelligence. His current research interests include biomedical signal processing, brain-computer interface, cognitive computing, and pattern recognition.

Feiping Nie (SM'21) received the PhD degree from Department of Automation, Tsinghua University, Beijing, China, in 2009. He is currently a full professor with School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, China. His research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval. He has published more than 100 papers in the following top journals such as TPAMI, IJCV, TIP, TNNLS/TNN, TKDE, Bioinformatics, and conferences such as ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, and ACM MM. His papers have been cited more than 10000 times. He is currently an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems, Information Fusion, Information Sciences, Pattern Recognition and Neural Processing Letters and a PC member for several prestigious such as AAAI and IJCAI.

Bao-Liang Lu (F'21) received the BS degree in instrument and control engineering from Qingdao University of Science and Technology, Qingdao, China, in 1982, the MS degree in computer science and technology from Northwestern Polytechnical University, Xian, China, in 1989, and the Dr.Eng. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1994. He was with Qingdao University of Science and Technology from 1982 to 1986. From 1994 to 1999, he was a Frontier Researcher with Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research (RIKEN), Nagoya, Japan, and a Research Scientist with the RIKEN Brain Science Institute, Wako, Japan, from 1999 to 2002. Since 2002, he has been a full professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include brain-like computing, neural networks, machine learning, brainCcomputer interaction, and affective computing. Prof. Lu received the IEEE Transactions on Autonomous Mental Development Outstanding Paper Award in 2018. He is an Associate Editor of IEEE Transactions on Affective Computing, IEEE Transactions on Cognitive and Developmental Systems and Journal of Neural Engineering.

Andrzej Cichocki (F'13) received the M.Sc. (hons.), Ph.D. and Dr.Sc. (Habilitation) degrees, in electrical engineering from Warsaw University of Technology, Warszawa, Poland, in 1972, 1976, and 1982, respectively. He spent several years at University Erlangen (Germany) as an Alexandervon-Humboldt Research Fellow and Guest Professor. He was a Senior Team Leader and Head of the laboratory for Advanced Brain Signal Processing, at RIKEN Brain Science Institute (Japan). Now he is a professor with Skolkov Institute of Science and Technology (Russia) and an adjunct professor with Hangzhou Dianzi University. He has authored more than 500 technical journal papers and 6 monographs in English (two of them translated to Chinese). His current research interests include multiway blind source separation, tensor decompositions, tensor networks, deep learning, human robot interactions and brain-computer interface. Dr. Cichocki served as an Associate Editor of, IEEE Transactions on Signal Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, Journal of Neuroscience Methods. He is the founding Editor-in-Chief for Computational Intelligence and Neuroscience Journal.