frontiers | Frontiers in Neuroscience

ORIGINAL RESEARCH
published: 27 April 2022
doi: 10.3389/fnins.2022.850906

Robust Latent Multi-Source Adaptation for Encephalogram-Based Emotion Recognition

Jianwen Tao[1†], Yufang Dan[1†], Di Zhou[2†] and Songsong He[1†]*

[1] *Institute of Artificial Intelligence Application, Ningbo Polytechnic, Ningbo, China,* [2] *Industrial Technological Institute of Intelligent Manufacturing, Sichuan University of Arts and Science, Dazhou, China*

In practical encephalogram (EEG)-based machine learning, different subjects can be represented by many different EEG patterns, which would, in some extent, degrade the performance of extant subject-independent classifiers obtained from cross-subjects datasets. To this end, in this paper, we present a robust Latent Multi-source Adaptation (LMA) framework for cross-subject/dataset emotion recognition with EEG signals by uncovering multiple domain-invariant latent subspaces. Specifically, by jointly aligning the statistical and semantic distribution discrepancies between each source and target pair, multiple domain-invariant classifiers can be trained collaboratively in a unified framework. This framework can fully utilize the correlated knowledge among multiple sources with a novel low-rank regularization term. Comprehensive experiments on DEAP and SEED datasets demonstrate the superior or comparable performance of LMA with the state of the art in the EEG-based emotion recognition.

Keywords: encephalogram, latent space, emotion recognition, co-adaptation, maximum mean discrepancy

OPEN ACCESS

Edited by:
Yuanpeng Zhang, Nantong University, China

Reviewed by:
Jian Zhang, Wuhan University of Technology, China
Fangfang Duan, Wuhan University of Technology, China

****Correspondence:***
Di Zhou
sion2005@sasu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *08 January 2022*
Accepted: *11 February 2022*
Published: *27 April 2022*

Citation:
Tao J, Dan Y, Zhou D and He S (2022) Robust Latent Multi-Source Adaptation for Encephalogram-Based Emotion Recognition. Front. Neurosci. 16:850906. doi: 10.3389/fnins.2022.850906

INTRODUCTION

Contemporarily, in the field of affective computing research, automated emotion recognition (AER) has attracted lots of attention from machine learning and computer vision (Kim et al., 2013). In traditional schema, one auto emotion recognition system driven by EEG signals usually includes two core components, i.e., feature extraction followed by emotion classification (Lan et al., 2018). Some representative EEG feature extraction methods (Jenke et al., 2014; Zhang et al., 2020b) can be viewed comprehensively in Jenke et al. (2014). This work mainly focuses on machine learning-based emotion classification methods.

In the past decade, a large scale of emotion recognition methods has been presented for effective emotion recognition using EEG features (Musha et al., 1997; Kim et al., 2013; Li et al., 2018b,c). Zheng (2017) proposed a novel emotion recognition method by exploiting the group sparse canonical correlation analysis, thus simultaneously implementing EEG channel selection and emotion recognition. Recently, Li et al. (2018c) also presented a sparse linear regression model with graph regularization for emotion recognition using EEG signals. In the past decade, due to their outperformed performance compared with traditional methods, deep emotion recognition methods using EEG signals have been widely explored in emotion feature extraction and recognition (Lotfi and Akbarzadeh-T, 2014), such as criminal psychological emotion recognition based on deep learning and EEG

signals (Liu and Liu, 2021), EEG-based Deep Belief Network model (Zheng and Lu, 2015), multi-channel EEG-based recognition model (Song et al., 2018), and EEG-based neural network model (Li et al., 2018b).

It is worthy to note that the aforementioned works for emotion recognition perform well only in such scenario that both training and test samples follow the same distribution (Zhang et al., 2020a), in which the recognition models obtained from the source dataset(s) therefore can be easily utilized in the target dataset effectively (Zhang et al., 2019a). Unfortunately, these traditional methods may fail in addressing cross-subject/dataset emotion recognition due to the mismatch of feature distribution with EEG signals. To address this issue, many domain adaptation (DA) emotion recognition models for AER problem have been promoted (Chu et al., 2017; Li et al., 2018a; Li et al., 2020a,b; Bao et al., 2021; Wang et al., 2021). In a DA emotion recognition system, one usually focuses on exploring an effective recognition model on one target domain with few or even none of the labeled data, by borrowing some positive knowledge from other source domain(s) with slightly different distribution with that of the target domain (Bruzzone and Marconcini, 2010; Tao et al., 2012; Long et al., 2014; Zhang et al., 2019b).

A typical challenge in one EEG-based emotion recognition system is the cross-subject/dataset learning problem (Li et al., 2018a). In such scenario, DA techniques can be exploited to address this challenging issue where both training and test data follow slightly different distribution (Tao et al., 2012; Long et al., 2014; Li et al., 2021). To deal with the challenging cross-subject EEG emotion recognition problem, Pandey and Seeja (2019) proposed a subject-independent approach for EEG emotion recognition. Li et al. (2018a) proposed another method for cross-subject EEG emotion recognition. In the past decade, deep neural networks (DNNs) (Ganin et al., 2016; Li et al., 2018a) have also driven rapid progress in DA (Duan et al., 2012a; Ding et al., 2018a). The DA issues can be solved by the domain adversarial neural network (DANN) (Ganin et al., 2016). It remains unclear, however, whether the performance of deep DA methods is really contributed by their deep feature representation, the fine-tuned classifiers, or is rather an outcome of the adaptation regularization terms (Ghifary et al., 2017).

Although existing DA methods have obvious effectiveness and efficiency in the special use of emotion recognition (Chu et al., 2017), there is few work to use the joint feature selection method and then carry out the multi-source adaptive domain recognition of cross datasets by exploiting the correlation knowledge among domains and features. Besides, during DA, most of the multi-source domain adaptation (MDA) methods (Yang et al., 2007; Duan et al., 2012b,c; Tommasi et al., 2014; Tao et al., 2015, 2017; Ding et al., 2018b) generally cope with the sources independently without considering the correlation information among the source domains (Zhang et al., 2019c), which may destroy the discriminant structure (either intrinsic or extrinsic) of multi-source domains (Rosenstein et al., 2005). Last but not the least, for an MDA system, it is crucial for source weight determination during learning based on the correlation and quality of source domains. To the best of our knowledge, these characters are not feasible enough in extant MDA methods.

In order to solve the above problems in existing MDA, we explore to exploit the relevant knowledge among sources in the uncovered subspaces to learn a multi-source adaptive emotion recognition model. In other words, we mainly adopt the strategy of digging the relationship between multi-source domains and one target domain (including feature and distribution) for promoting multi-source adaptive emotion recognition with EEG signals. We aim to progress beyond existing works that have partially addressed those issues by exploring to solve all the above-mentioned issues in a unified framework. Specifically, we propose in this work a robust Latent Multiple-source Adaption (LMA) method for EEG-based emotion recognition by mining multiple shared latent subspaces, each for one source–target domain pair. The method employs the robust regression scheme to process high-dimensional, sparse outliers and non-i.i.d. (independently identical distribution) EEG features by jointly utilizing the $l_{2,1}$-norm (Nie et al., 2010a) and trace norm. Under this framework, the row sparsity regularization is designed to obtain the solution of sparse feature selection (Zhang et al., 2020b). We match distributions between each domain pair (including both target and multi-source domains) by minimizing the nonparametric Maximum Mean Discrepancy (MMD) (Gretton et al., 2009; Pan et al., 2011) in each uncovered latent space shared by this source–target pair. The contributions of this paper are listed as follows:

(1) We propose a unified multi-source adaptive emotion recognition framework with EEG features by uncovering multiple latent subspaces.
(2) Our framework selects features in a collaborative way and considers the correlated knowledge among sources. In LMA, the importance of each feature does not need to be evaluated separately. In addition, in our unified framework, we can learn multiple loss functions with feature selection for all source adaptation subjects synchronously, so that our framework can use the correlated information of multiple sources as auxiliary information.
(3) In this framework, the original geometric structure is retained by using the graph Laplacian regularization, and the $l_{2,1}$-norm minimization sparse regression approach is used to suppress the influence of noise or outliers in the domains, which shows the robustness of the framework.
(4) Through a large number of experiments on two EEG datasets, we prove the effectiveness and convergence of this framework.

The remainder of the paper is organized as follows: In section Related Work, we discussed the related works with feature selection and multi-source DA learning. In section Proposed Framework, our framework LMA will be designed, and section Algorithm arranges the corresponding optimal algorithm of LMA. The experimental results and analysis on two real EEG datasets are presented in section Experimental Evaluation. Finally, we conclude in Section Conclusion.

TABLE 1 | Notations and descriptions.

Notations	Descriptions
N	Sample number of each source–target pair
d	Feature dimensionality number
χ	Sample/feature space
Γ	Label/prediction space
$a = [a_1, a_2, ..., a_d]^T \in \mathbb{R}^d$	Vector a
$A \in \mathbb{R}^{n \times d}$	Matrix A
$A_{i,j}$	The (i, j)th element of the matrix A
$A_{i,:}$ and $A_{:,j}$	The i/jth row/column vector of A
$(\cdot)^T$	Transpose operator
$tr(\cdot)$	Trace operator
$\langle A, B \rangle = tr(A^T B)$	The inner product of two matrices A and B
$\|\|a\|\|_p := \left(\sum_{i=1}^{d} \|a_i\|^p\right)^{1/p}$	The p-norm of a vector a
$\|\|A\|\|_{2,1} = \sum_{i=1}^{n} \|\|A_{i,:}\|\|_2 = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{d} A_{ij}^2}$	The $l_{2,1}$-norm of A
$\|\|A\|\|_* = tr((AA^T)^{\frac{1}{2}})$	The trace-norm of A
I_r	Identity matrix of size $r \times r$
1_d	d-dimensional vector with all ones
0_d	d-dimensional vector with all zeroes

RELATED WORK

In the past decades, affective computing community has paid increasing attention to the emotion recognition with brain–computer interfaces (BCI) (Mühl et al., 2014; Chu et al., 2017). A brain–computer interface system could capture certain emotion states and respectively make corresponding response to these states using spontaneous EEG signals even when explicit input from the subjects is unavailable (Zhang et al., 2019a), thus augmenting the user experience in the session of interactivity. Nowadays, a large number of methods (Zhang et al., 2016, 2017) have been proposed to recognize different emotion information from brain-wave signals. The latest works about affective BCI (aBCI) took account of machine learning algorithms on emotion recognition using a few discriminative features (Jenke et al., 2014; Mühl et al., 2014). In one representative BCI system, a certain feature extractor firstly extracts discriminative features from the raw EEG data, and then these features as well as labeled emotion states are sent into the classifier for real-time affection recognition. In the last decade, many aBCI-related works have presented sound and interesting emotion recognition performance (Mühl et al., 2014).

Although existing methods have obtained satisfied achievements on EEG-based emotion recognition, the expected performance could still be degraded by certain impacts in the case of cross-subject/dataset recognition due to the difference between subjects/datasets. Therefore, one needs to train a specific classifier for individual subject/dataset-of-interest. Even for the same subject, it is also indispensable to recalibrate the classifier frequently for maintaining a satisfied recognition accuracy since the EEG signals are unstable now and then. This would undoubtedly increase the costs of manual labor as well as time. Fortunately, the DA (a.k.a. domain transfer) technique can be leveraged to tackle these issues existing in EEG-based emotion recognition.

In the past decade, DA technique (Duan et al., 2012a,b,c; Tzeng et al., 2015; Tzeng et al., 2017; Ding et al., 2018a,b,c) has elicited an increasing attention in the community of machine learning. Up to now, domain-adaptation-based emotion recognition methods have nearly dominated the literature of aBCI Dolan, 2002; Mühl et al., 2014; Jayaram et al., 2016; Lan et al., 2018; Zhong et al., 2020; Zheng et al., 2015; Zheng and Lu, 2016; Chai et al., 2016; Chai et al., 2017; Shi et al., 2013; Koelstra et al., 2012; Zheng and Lu, 2015), which aim to address different issues in emotion classification by pursuing various DA skills using the EEG datasets such as SEED. In these preceding works, a commonly used strategy is to uncover a shared subspace from different domains by preserving certain discriminative properties, thus decreasing the differences among subjects or sessions extracted from the captured EEG signals (Tao and Dan, 2021). While extensive exploration on cross-subject/session has been conducted effectively in the prior works by leveraging various DA tricks, one obvious shortage in these works is that the evaluation dataset is just limited to one single database, e.g., SEED. In practical aBCI applications, the EEG datasets could also change since the EEG signals may be produced by different subjects, sessions, EEG devices, experimental schemes, and emotional stimuli. Henceforth, one of the yet unsolved issues in current research is the robustness and effectiveness of the proposed DA methods on cross-datasets/subjects.

PROPOSED FRAMEWORK

Notation

In the context, the symbol definitions are listed in **Table 1**. We respectively denote by $[A_1, A_2, ..., A_k]$ and $[A_1; A_2; ...; A_k]$ the concatenation of k matrices according to the row (horizontally) and the column (vertically). In this work, we focus on the multi-source adaptation framework, which can be driven by S source domains of c-class. We denote by $X^a = \{x_1^a, ..., x_{n_a}^a\} \in \mathbb{R}^{d \times n_a}$ $(a = 1, 2, ..., S)$ the ath source dataset with n_a samples[1]. Its corresponding class label matrix can be denoted as $Y^a = [y_1^a, ..., y_{n_a}^a]^T \in \mathbb{R}^{n_a \times c} \in \Gamma = \{0, 1\}^{c \times 1}$ with $y_{il} = 1$ if the ith sample is labeled as the lth class and -1 others. Correspondingly, we denote by $X^t = \{x_1^t, x_2^t, ..., x_m^t\} \in \mathbb{R}^{d \times n_t}$ the target dataset of interest. Since the true classes of the samples in X^t are inaccessible in the training stage, the target labels (or pseudo labels) $Y^t = [y_1^t, ..., y_{n_t}^t]^T \in \mathbb{R}^{n_t \times c} \in \Gamma$ can be predicted by certain pre-trained classifiers trained on the source datasets with labeled data. Therefore, detecting the ground-truth label of each target sample is our ultimate goal.

We further denote X^a/X^t with the label $\bar{l}$ as $X^{a(\bar{l})}/X^{t(\bar{l})}$ $(\bar{l} = 1, ..., c)$, and the ath source–target domain pair as $X_a = [X^a, X^t] \in \mathbb{R}^{d \times N}$ $(N = n_t + n_a)$ with label matrix $Y_a = [Y^a, Y^t]$ by packing the ath source and the target data.

[1]While we do not need to limit the number of instances in each source domain, which is identical with that assumed when shaped into the training matrix, for the sake of simplicity, we can extract the same number of training instances from each source domain.

Problem Statement

A commonly used strategy in the representative MDA is to acquire knowledge from multiple sources by leveraging certain common knowledge shared by them to promote the target learning of interest. We propose in this work a robust Latent Multiple-source Adaption (LMA) emotion recognition method based on EEG features. The method employs the robust regression scheme to process high-dimensional, sparse, outliers, and non-i.i.d. EEG features by jointly utilizing the $l_{2,1}$-norm and trace norm (Yang et al., 2013). The designed method has three characteristics, which are integrated into a unified optimization formulation to find an effective emotion recognition model by aligning the feature distribution between each source–target domain pair. Specifically, it includes four technical aspects: (1) *via* employing the $l_{2,1}$-norm minimization, a robust loss term is introduced into each source model learning by taking account of the influence of noise or outliers in EEG signal (Li et al., 2015), and a sparse regularization term is designed to eliminate over-fitting and a sparse features subset is selected; (2) based on the designed regression model and the semantic distribution matching between each pair of domains in each uncovered latent spaces (Tao et al., 2019), it not only provides robustness on loss function, but also retains the domain distribution (including local and global) structures (Nie et al., 2010b), and meanwhile maintains a high dependence on the (pseudo) label knowledge of the source domains and the target domain (Nie et al., 2010b; Ding et al., 2018c; Zhang et al., 2020a), so as to obtain preferable generalization performance; and (3) by exploiting the trace norm of matrix, we can make full use of the correlative information among multiple sources and transfer more discriminative knowledge to the target domain.

Specifically, we present the flow diagram of LWA in **Figure 1** to illustrate our innovation: firstly, we can project each source EEG data into one domain-invariant subspace by minimizing the domain-wise distribution discrepancy; thus, S classifiers are being jointly learned by employing trace norm as well as $l_{2,1}$-norm; we then obtain S target label matrices predicted from these source classifiers on the target domain; furthermore, in the original space, we also learn a target model using the squared regression scheme with the constraint of prediction consistency on the target data between those source models and the target model; and by uncovering multiple domain-invariant latent spaces, we finally formulate a joint learning framework of multi-source adaptation for EEG-based emotion recognition. To implement these properties, in the following part, we will detail the objective formulation of the proposed method.

General Formulation

In this section, we propose the general formulation of LMA framework underpinned by the robust regression principle and the regularization theory. We investigate the learning problem under multiclass setting, with the decision classifiers $\left\{f_\theta^a(x)\right\}_{a=1}^S$, where θ is the parameter of the hypothesis space of those decision functions. We propose a unified MDA framework by uncovering S discriminative latent subspaces Θ_a $(a = 1, 2, ..., S)$ (Tao and Xu, 2019) and to learn decision classifiers $f_\theta^a(x)$s of all sources simultaneously. In particular, the proposed method minimizes the distribution difference of each domain pair after the projection P_a into the subspace Θ_a, as well as the structural risk functional of the labeled data from the source domain X^a. We also let Θ_a be orthogonal on rows so that $\Theta_a\Theta_a^T = I_{r\times r}$, where $r(\ll d)$ is the dimensionality of the shared latent space. We then endeavor to find S cross-domain models parameterized by $\{\theta_a\}_{a=1}^S$ *via* jointly utilizing correlated knowledge among sources in some latent spaces. We therefore propose the following general formulation of LMA:

$$\begin{aligned}&\Re\left(f_{\theta_a}^a, f_{\theta_t}^t, \Theta_a\right)\\&=\sum_{a=1}^{S}\left(R(f_{\theta_a}^a, Y_a)+dist_{\Theta_a}(X^a, X^t)+\mho(f_{\theta_a}^a)\right)+Con(f_{\theta_a}^a, f_{\theta_t}^t, X^t)\\&\qquad+R(f_{\theta_t}^t, Y^t)+\Omega(f_\theta^a),\end{aligned}\tag{1}$$

where $\theta_{a/t}$ is the parameter set of the ath source/target model, $Con(f_{\theta_a}^a, f_{\theta_t}^t, X^t)$ enforces the discriminative consistency between the ath source and the target model on the target dataset, $R(\cdot,\cdot)$ is the robust regression model, the regularization term $\mho(f_{\theta_a}^a)$ controls the complexity of $f_{\theta_a}^a$, $dist_{\Theta_a}(X^a, X^t)$ is for aligning the distributions of each domain pair in the latent space Θ_a, the regularizer $\Omega(f_\theta^a)$ controls the low rankness of all source models for mining the correlated knowledge. Hence, by solving the objective in (1), the subspace Θ_a and the decision functions $f_{\theta_{a/t}}^{a/t}$s can be learned simultaneously. In the sequence, we will focus on designing these components in the general formulation one by one to construct a unified framework.

Design of Regression Model

In LMA, we learn a composite source classifier function $f^a = P_a^\circ\Theta_a$ trained on the EEG features, where Q_a represents the source classifier model, and $\circ$ is the function combination operator. We therefore explore to find the best approximation W_t for f^t by leverage Q_a in Θ_a with the assumption that there exist some commonalities (e.g., discriminative structures) between different domains (Tao et al., 2019). Moreover, it should also maintain the discriminative structure in the original space. To capture the source correlation information, we respectively design the following classification functions for the ath source domain in the latent space:

$$f_\theta^a(X^a) = V_a^T\phi(X^a) + Q_a^T\psi_\theta(X^a),\quad a = 1, 2, ..., S,\tag{2}$$

where $\phi : \chi \to H^2$ is a known feature map projecting the ath source data from the input space χ into certain reproducing kernel Hilbert space (RKHS) (Nie et al., 2010b) H. The other component ψ_θ is a parameterized low-dimensional space that aims at encoding the shared structure between each source domain and the target domain. The weight vector Q_a is defined in the projected subspace under the projection ψ_θ for the ath source, and V_a is the weight matrix defined in the original feature space. With the parametric form in (2), the learned subspace

[2]It is important to note that the feature mapping function ϕ with respect to each source domain can be completely different from each other.

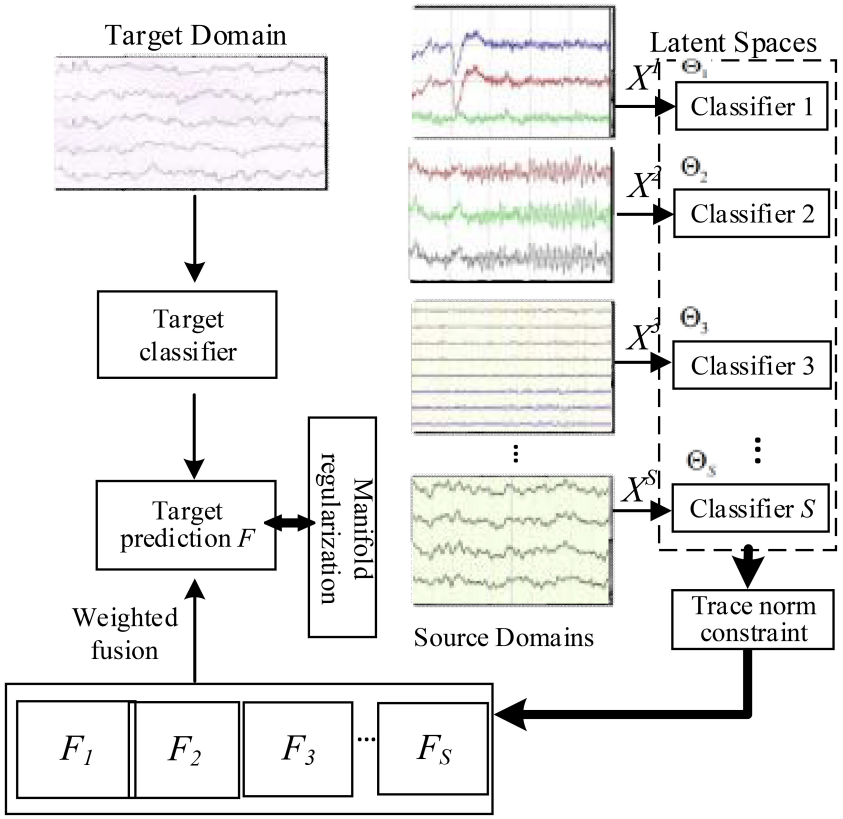

FIGURE 1 | Flowchart of LMA on EEG-based emotion recognition.

ψ_θ can capture the intrinsic structure of source correlation in the MDA problem, which are shared by each source and the target domain. Correspondingly, we also can design the target recognition model: $f^t(X^t) = W_t^T \phi(X^t)$ with the weight matrix W_t. We present the empirical kernel map as discussed in Gretton et al. (2009):

$$\begin{aligned} &\psi_e : \quad \chi \to R^N, \ \textit{for linear kernel mapping} \\ &x \to K_\psi(\cdot, x)|_{x_1,x_2,\ldots,x_N} = \left(K_\psi(x_1, x), \ldots, K_\psi(x_N, x)\right), \ . \\ &\textit{for nonlinear kernel mapping} \end{aligned}$$

In the following, we will discuss both the linear classification function and the nonlinear (kernel) classification function and integrate them into a unified form.

- *Linear classifier*. We can consider a simple linear form of feature map, where $\theta = \Theta_a$ is an $r \times d$ dimensional matrix and $\psi_\theta(x) = \Theta_a \psi(x)$, with a known d-dimensional vector function $\psi(x)$. Furthermore, following (11), we can take a simple model $\phi(X^a) = \psi(X^a) = X^a$ into account. We can thereby write the linear classifier as

$$f_\theta^a(X^a) = V_a^T X^a + Q_a^T \Theta_a X^a, \quad a = 1, 2, \ldots, S. \tag{3}$$

- *Nonlinear classifier*. If we take kernel learning into account and assume that the feature map $\phi(x)$ and $\psi(x)$ belong to certain reproducing kernel Hilbert space (RKHS), Eq. (3) therefore can be kernelized. For $\psi(x)$, we firstly denote the kernel matrix as $K_\psi = \langle \psi(x_i), \psi(x_j) \rangle$. By using empirical kernel, we have kernel matrix $K^a = \phi(X^a)$ with $(K^a)_{i,j} = \left\langle \phi(x_i^a), \phi(x_j^a) \right\rangle$, where $x_i^a, x_j^a \in X^a$. Finally, we can let $\psi_\theta = \Theta_a \psi_e$, where $\Theta_a \in R^{r \times N}$ is used to transform the empirical kernel vector to an r-dimensional space. Let $\{\Psi_a^i\}_{i=1}^r$ denote the weight parameters in the embedded kernel subspace for the ath source. Hence, the kernelized decision functions become

$$f_\theta^a(x) = \omega_a^T K^a(\cdot, x) + \Psi_a^T \Theta_a K^a(\cdot, x). \tag{4}$$

where ω_a is the weight coefficients in the original kernel space for the ath source.

In order to model the linear case in Eq. (3) and kernel case in Eq. (4) into a unified framework, we introduce

$$T_a = \begin{cases} V_a + Q_a^T \Theta_a, & \textit{linear} \\ \omega_a + \Psi_a^T \Theta_a & \textit{kernel} \end{cases}. \tag{5}$$

Moreover, in the following, we use two symbols, namely, W_a and P_a, where W_a denotes V_a in the linear case and ω_a in the kernel case, and P_a denotes Q_a in the linear case and Ψ_a in the kernel case. Then, Eq. (5) becomes $T_a = W_a + P_a^T \Theta_a$ for both linear and

kernel cases, and we can represent the data in linear space and nonlinear space as follows:

$$\overline{X}^{a/t} = \begin{cases} X^{a/t}, \ linear \\ K^{a/t}(\cdot, x) \ , \ kernel \end{cases}. \tag{6}$$

In the sequence, we also refer to $X^{a/t}$ as $\overline{X}^{a/t}$ if without special denotation for simplicity of expression. As a result, we can formulate the predictors, linear form as in Eq. (4) and nonlinear form as in Eq. (6), in a unified form as depicted in

$$f_\theta^a(X^a) = T_a^T X^a, \quad a = 1, 2, ..., S \tag{7}$$

We introduce the sparse regression scheme (Shi et al., 2015) by exploiting $l_{2,1}$-norm minimization to enhance the robustness against the misclassification. We particularly construct a scaled pseudo label matrix for the target data, i.e., $F = [f_1, f_2, ..., f_{n_t}] = (Y^t(Y^t)^T)^{-1/2}Y^t \in \mathbb{R}^{n_t \times c}$, where the scaled pseudo label $f_i = y_i^t$ if x_i^t is labeled, $f_i = 0$ otherwise. Therefore, $FF^T = I_{n_t}$ can be easily derived with additional constraint $F \geq 0$. We then respectively find the source classifiers trained on X^a ($a = 1, 2, ..., S$) and the target classifier trained on X^t by minimizing the following loss functions.

$$R(f_{\theta_a}^a, Y_a) = \left|\left| f_{\theta_a}^a(X^a) - Y^a \right|\right|_{2,1}. \tag{8}$$

$$R(f^t, F) = \left|\left| f^t(X^t) - F \right|\right|_F^2 + \beta \left(||W_t||_{2,1} + tr(FLF^T) \right). \tag{9}$$

where L denotes the graph Laplacian matrix induced from the target samples.

Moreover, it is intuitively reasonable that the outputs of f^a s on the target domain are expected to be consistent with those of f^t, which would gradually make P_a and W_t more accurate after lots of iterations. This prediction consistency can be minimized *via* the following residual:

$$\begin{aligned} Con(f^a, f^t, X^t) = & \left|\left| X_t^T(W_a + \Theta_a P_a) - F^a \right|\right|_F^2 \\ & + \left|\left| X_t^T W_t - F \right|\right|_F^2 + \left|\left| F^a - F \right|\right|_F^2. \end{aligned} \tag{10}$$

In such a way, P_a and W_t would jointly enhance the target discriminations for the final emotion recognition.

Additionally, based on the parametric form of the decision function f_θ^a as in Eq. (7), we introduce the following regularizer:

$$\mho(f_\theta^a) = ||W_a||^2 + ||T_a||_{2,1} = \left|\left| T_a - \Theta_a^T P_a \right|\right|_F^2 + ||T_a||_{2,1}, \tag{11}$$

which controls the complexity of each source classifier independently in the original and latent subspaces, respectively.

Uncovering Latent Spaces

In this subsection, we will present an effective strategy to capture multiple domain-invariant subspaces to mitigate the domain discrepancy as well as excavate some domain-invariant discriminative information. To this end, we give two main constraints or conditions on uncovering these latent spaces: (1) preserving the within-domain local structures and (2) aligning the inter-domain marginal distribution divergence as well as conditional distribution discrepancy. Following existing feature extraction methods (Tao et al., 2016), we further constrain Θ_a to be orthogonal on rows, i.e., $\Theta_a \Theta_a^T = I_r$, where r (typically far less than d) is the feature dimensionality in the latent space.

To fulfill the first condition, we construct a locality preserving regularizer to measure the smoothness along the intrinsic discriminative structure of the domain features (Nie et al., 2010b; Shi et al., 2015; Ding et al., 2018c). Specifically, one can construct an undirected graph with a weighted adjacency matrix $\prod_a = [(\prod_a)_{i,j}]_{i,j=1,2,...,N}$, which is defined as (Yan et al., 2006):

$$(\prod_a)_{i,j} = \begin{cases} \exp(-\gamma||x_i - x_j||^2), \ \text{if } x_i \in \delta_k(x_j) \text{ or } x_j \in \delta_k(x_i) \\ \quad \text{and both have the same labels} \\ \exp(-\frac{\gamma}{||x_i - x_j||^2}), \ \text{if } x_i \in \delta_k(x_j) \text{ or } x_j \in \delta_k(x_i) \\ \quad \text{and both have different labels,} \\ 0, \ \text{otherwise} \end{cases}, \tag{12}$$

where $x_i, x_j \in X_a = [X^a, X^t]$, $\delta_k(x)$ denotes the k nearest neighbor set of x, and the hyper-parameter γ can be empirically computed as $\overline{\theta}_a\sqrt{c}$ due to the impact of multi-class distribution, where $\overline{\theta}_a$ is the squared root of the mean norm of X_a. Deriving a diagonal matrix Δ_a from $\prod_a$ with $(\Delta_a)_{i,i} = \sum_j (\prod_a)_{i,j}$, we then compute the graph Laplacian matrix as $L_a = \Delta_a - \prod_a$. Thus, preserving the local geometrical structures of EEG features can be implemented by the following commonly used formulation in the manifold learning (Chen et al., 2013).

$$tr(\Theta_a^T X_a L_a X_a^T \Theta_a). \tag{13}$$

Benefiting from its simplicity and effectiveness, Maximum Mean Discrepancy (Gretton et al., 2009; Pan et al., 2011) has been commonly used to measure the distribution distance between two different domains. Consequently, to meet the second condition, we aim to minimize the MMD in certain optimized RKHS (Gretton et al., 2009). Specifically, the MMD between each domain pair is defined as follows:

$$\begin{aligned} & MMD(X^a, X^t) \\ & = \sup_{||\phi|| \leq 1} \left(E_{X^a \sim P}[\phi(X^a)] - E_{X^t \sim Q}\left[\phi(X^t)\right] \right), \\ & = \left|\left| E_{X^a \sim P}[\phi(X^a)] - E_{X^t \sim Q}\left[\phi(X^t)\right] \right|\right|_H \end{aligned} \tag{14}$$

The empirical counterpart of the MMD in Eq. (14) can be defined as:

$$MMD(X^a, X^t) = \left|\left| \frac{1}{n_a} \sum_{x_i \in X^a} \phi(x_i) - \frac{1}{n_t} \sum_{x_j \in X^t} \phi(x_j) \right|\right|_H, \tag{15}$$

which can recover an asymptotically unbiased estimation of the squared MMD in Eq. (14). Denote the gram matrix $\widetilde{K}_a(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ on dataset X_a as

$$\widetilde{K}_a = \begin{bmatrix} \widetilde{K}^a & \widetilde{K}^{at} \\ \widetilde{K}^{ta} & \widetilde{K}^t \end{bmatrix} \in R^{N \times N}, \tag{16}$$

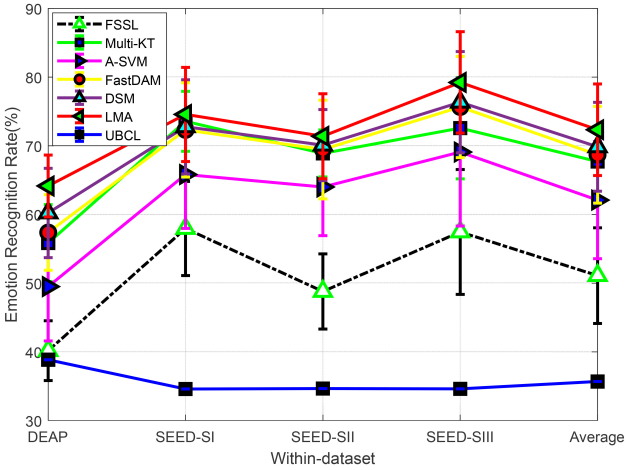

FIGURE 2 | Domain adaptation emotion recognition on within-dataset (SI: Session I, SII: Session II, SIII: Session III).

where $\widetilde{K}^a$, $\widetilde{K}^t$, and $\widetilde{K}^{at}$ (or $\widetilde{K}^{ta}$) are the Gram matrices respectively defined on the source domain, target domain, and cross domain data. Thus, the squared MMD in Eq. (15) can be formulated as

$$MMD(X^a, X^t) = tr\left(\widetilde{K}_a D^a\right), \tag{17}$$

where

$$D_{ij}{}^a = \begin{cases} \frac{1}{n_a^2} & when\ x_i, x_j \in X^a \\ \frac{1}{n_t^2} & when\ x_i, x_j \in X^t \\ \frac{-1}{n_a n_t} & otherwise \end{cases}. \tag{18}$$

In the sequence, we will take into account the feature map ϕ in linear as well as kernel forms:

λ Linear kernel: $\widetilde{K}_a = X_a^T \Theta_a^T \Theta_a X_a$ if $\phi(x) = \Theta_a x$, where $X_a = [x_1, x_2, ..., x_N]$.

λ Nonlinear kernel: $\widetilde{K}_a = K_a^T \Theta_a^T \Theta_a K_a$ if $\phi(x) = \Theta_a \psi_e(x) = \Theta_a K_a(\cdot, x)$, where $\psi_e(x)$ is the empirical kernel map defined in Eq. (4).

Recalling the definition of $\overline{X}$ in Eq. (6), the domain discrepancy criterion defined in (17) can be reformulated:

$$MMD^2(X^a, X^t) = tr(\widetilde{K}_a D^a) = tr(\Theta_a^T \overline{K}_a \Theta_a), \tag{19}$$

where

$$\overline{K}_a = \begin{cases} X_a D^a X_a^T, & linear\ kernel \\ K_a^T D^a K_a, & nonlinear\ kernel \end{cases}. \tag{20}$$

Note that Eq. (19) could not preserve the local structures of the EEG data from the same class in the latent spaces due to the shortage of semantic alignment. This would significantly deteriorate the learning performance in some cases. To this end, we further address this issue by improving Eq. (19) with the following class distribution matching term:

$$\sum_{l=1}^{c} \left\| \frac{1}{n_a^l} \sum_{i=1}^{n_a^l} \Theta_a^T \phi(x_i^{a(c)}) - \frac{1}{n_t^l} \sum_{j=1}^{n_t^l} \Theta^T \phi(x_j^{t(c)}) \right\|_F^2$$

$$= \sum_{l=1}^{c} tr(\Theta_a^T \overline{K}_a^{(l)} \Theta_a), \tag{21}$$

where

$$\overline{K}_a^l = \begin{cases} X_a^{(l)} D^{a(l)} (X_a^{(l)})^T, & linear \\ (K_a^{(l)})^T D^{a(l)} K_a^{(l)}, & kernel \end{cases},$$

$X_a^{(l)} = [X^{a(l)}, X^{t(l)}]$, with $X^{a(l)}$ and $X^{t(l)}$ being the datasets of the lth class, respectively, from source and target domains, n_a^l (respectively n_t^l) is the data size of the lth class from the ath source (respectively target) domain, and the elements of the matrix $D^{a(l)} = [Di, j^{a(l)}]$ are defined as

$$D_{i,j}^{a(l)} = \begin{cases} \frac{1}{(n_a^l)^2}, & when\ x_i, x_j \in X^{a(l)} \\ \frac{1}{(n_t^l)^2}, & when\ x_i, x_j \in X^{t(l)} \\ \frac{-1}{n_a^l n_t^l}, & otherwise \end{cases}. \tag{22}$$

Equation (21) explicitly forces EEG data from different domains but the same class to be mapped adjacently in the latent spaces. By unifying Eq. (19) and Eq. (21), we can obtain:

$$\sum_{l=0}^{c} tr(\Theta_a^T \overline{K}_a^{(l)} \Theta_a) = tr\left(\Theta_a^T \left(\sum_{l=0}^{c} \overline{K}_a^{(l)}\right) \Theta_a\right), \tag{23}$$

where $\overline{K}a^{(0)} = \overline{K}a$ and $D^{a(0)} = D^a$. We further denote $\Lambda^a = \sum_{l=0}^{c} \overline{K}_a^{(l)}$. By combining Eq. (13) and Eq. (23), we can attempt to uncover a latent space by minimizing the following formulation:

$$dist_{\Theta_a}(X^a, X^t) = tr\left(\Theta_a^T (X_a L_a X_a^T + \Lambda^a) \Theta_a\right)\ \ s.t.\ \Theta_a^T \Theta_a = I_r. \tag{24}$$

Sharing Source Discriminative Structure

While each source model is learned in different latent space from each other, one still can presume that these source models might be correlated due to the correlation of source EEG signals in the

model level (Tao et al., 2016). These correlated discriminative structures can be encoded by a low rank matrix of all source models, thus transferring the source knowledge from each other. For its simplicity of computation, the following trace norm of the matrix $P = [P_1, P_2, ..., P_S]$, which is a surrogate of the rank minimization, can be adopted for correlating the source models.

$$\Omega(f_\theta^a) = ||P||_* = tr\left(\left(PP^T\right)^{\frac{1}{2}}\right), \tag{25}$$

Overall Formulation

Combining the above formulations respectively defined in Eqs (8) to (11) and Eqs (24) and (25) together, we have the following objective function:

$$\begin{aligned}\Re = \underset{\substack{P_a, W^t, T_a, F_a,\\ F, W^t, \vartheta, \eta, \Theta}}{\arg\min} & \sum_{a=1}^{S}\left[\vartheta_a^{q_1}\left|Y_a - X_a^T T_a\right|_{2,1}\right. \\ & \left.+\alpha\left(|T_a|_{2,1} + |T_a - \Theta_a P_a|_F^2\right) + \eta_a^{q_2}\left|F_a - X_t^T T_a\right|_{2,1}^2\right], \\ & + \sum_{a=1}^{S}\eta_a^{q_2} tr(\Theta_a^T C_a \Theta_a) + \sum_{a=1}^{S}\eta_a^{q_2}|F_a - F|_F^2 \\ & + \left|X_t^T W_t - F\right|_F^2 + \beta\left(|W_t|_{2,1} + tr(FLF^T)\right) + \frac{\lambda}{2}|P|_* \\ s.t.\ & \sum_{a=1}^{S}\vartheta_a = \sum_{a=1}^{S}\eta_a = 1, \Theta_a^T\Theta_a = I_r\end{aligned} \tag{26}$$

where $C_a = X_a L_a X_a^T + \Lambda^a$, $\vartheta = [\vartheta_1, ..., \vartheta_a]^T$ is the weight vector to jointly combine all source regression loss, α, β, and λ are three regularization parameters, $q_1, q_2 > 1$ are two tunable parameters for avoiding trivial solution, and the tunable vector $\eta = [\eta_1, \eta_2, ..., \eta_S]$ denotes the adaptation degrees of different sources. The $l_{2,1}$-norm regularization added on projection matrix P_a forces most of the rows in P_a ($a = 1, ..., S$) to shrink to zero, thus performing feature selection on original data.

Emotion Recognition

After the best model parameters have been pursued, the source and target classifiers can be applied to recognize the emotion level of each probe EEG data. Specifically, we linearly fuse two recognition results on the probe data, i.e., $f^s(X^t) = \sum_{a=1}^{S}\vartheta_a (X^t)^T(\Theta_a P_a + W_a)$ obtained from source models, and $f^t(X^t) = (X^t)^T W^t$ predicted from the target model, as the final prediction value. That is, the following combination function can be exploited for recognizing the emotion level of the given test data x_i^t:

$$j = \arg\max_j \left(y_i^t = \delta f^s(x_i^t) + (1-\delta) f^t(x_i^t)\right)_j,$$

where δ is a trade-off parameter, tuned from $[0, 1]$. In the experimental setting, we empirically set $\delta = 0.5$ for initialization, followed by the evaluation of its impact on performance with different values of it.

ALGORITHM

In this section, an alternately iterative procedure is adopted to optimize the objective function in Eq. (26), which is followed by an overall algorithm.

Optimization

In terms of the definition in Nie et al. (2010a), we can derive $\left|\widetilde{T}\right|_{2,1} = 2tr\left(\widetilde{T}^T Q \widetilde{T}\right)$, where Q is a diagonal matrix with the ith diagonal element $Q_{ii} = \frac{1}{2||\widetilde{T}_{i,:}||_2}$. Hence, we can further transform Eq. (26) into Eq. (27):

$$\begin{aligned}\Re = \underset{\substack{P_a, W_t, T_a, F_a,\\ F, \vartheta, \eta, \Theta}}{\arg\min} & \sum_{a=1}^{S}\begin{pmatrix}\vartheta_a^{q_1} tr(\widetilde{T}_a^T Z_a \widetilde{T}_a) + \alpha tr\left(T_a^T G_a T_a\right. \\ \left.+|T_a - \Theta_a P_a|_F^2\right) \\ +\eta_a^{q_2}\left(tr(\Theta_a^T C_a \Theta_a) + |F_a - F|_F^2\right. \\ \left.+tr(Q_a^T \widetilde{Z}_a Q_a)\right)\end{pmatrix}, \\ & + \frac{\lambda}{2} tr\left(P^T\left(PP^T\right)^{-\frac{1}{2}}P\right) + \left|X_t^T W_t - F\right|_F^2 + \beta\left(|W_t|_{2,1} + tr(FLF^T)\right) \\ s.t.\ & \sum_{a=1}^{S}\vartheta_a = \sum_{a=1}^{S}\eta_a = 1, \Theta_a^T\Theta_a = I_r\end{aligned} \tag{27}$$

where $\widetilde{T}_a = (Y_a - X_a^T T_a)$ and $Q_a = F_a - X_t^T T_a$.

By solving the derivative of Eq. (27) w.r.t. W_t and letting it equal to zero:

$$\frac{\partial \Re}{\partial W_t} = 0 \Rightarrow W_t = \Delta^{-1} X_t F, \tag{28}$$

where $\Delta = X_t X_t^T + \beta\widetilde{V}$, and $\widetilde{V}$ is a diagonal matrix with the kth diagonal element being $(\widetilde{V})_{kk} = \frac{1}{2||(W_t)_{k,:}||_2}$. Substituting W_t in Eq. (27) with Eq. (28), we have:

$$\begin{aligned}\Re = \underset{\substack{P_a, T_a, F_a,\\ F, \vartheta, \eta, \Theta}}{\arg\min} & \sum_{a=1}^{S}\begin{pmatrix}\vartheta_a^{q_1} tr(\widetilde{T}_a^T Z_a \widetilde{T}_a) + \alpha\left(tr\left(T_a^T G_a T_a\right)\right. \\ \left.+|T_a - \Theta_a P_a|_F^2\right) + |F_a - F|_F^2 \\ +\eta_a^{q_2}\left(tr(\Theta_a^T C_a \Theta_a) + tr((F_a - X_t^T T_a)\right. \\ \left.^T \widetilde{Z}_a (F_a - X_t^T T_a))\right)\end{pmatrix}, \\ & + \frac{\lambda}{2} tr\left(P^T\left(PP^T\right)^{-\frac{1}{2}}P\right) + \beta tr\left(F\widetilde{\Delta}F^T\right) \\ s.t.\ & \textstyle\sum_{a=1}^{S}\vartheta_a = \sum_{a=1}^{S}\eta_a = 1, \Theta_a^T\Theta_a = I_r\end{aligned} \tag{29}$$

where $\widetilde{\Delta} = \beta L + \left(X_t^T \Delta^{-1} X_t - I\right) + \beta X_t^T \Delta^{-1} \widetilde{V} \Delta^{-1} X_t$. By solving the derivative of Eq. (29) w.r.t. F_a and letting it equal to zero, we have:

$$\begin{aligned}\frac{\partial \Re}{\partial F_a} &= 2\widetilde{Z}_a F_a - 2\widetilde{Z}_a X_t^T T_a + 2F_a - 2F = 0 \\ &\Rightarrow F_a = S_a^{-1}\left(F + \widetilde{Z}_a X_t^T T_a\right)\end{aligned}, \tag{30}$$

where $S_a = \widetilde{Z}_a + I$. Plugging F_a in Eq. (30) into Eq. (29), we can get:

$$\begin{aligned}&\eta_a^r tr\left(\left(S_a^{-1}F + \left(S_a^{-1}\widetilde{Z}_a X_t^T - X_t^T\right)T_a\right)^T \widetilde{Z}_a \left(S_a^{-1}F + \left(S_a^{-1}\widetilde{Z}_a X_t^T - X_t^T\right)T_a\right)\right) \\ &+\eta_a^r \left|\left|(S_a^{-1} - I)F + S_a^{-1}\widetilde{Z}_a X_t^T T_a\right|\right|_F^2 + \beta tr\left(F\widetilde{\Delta}F^T\right)\end{aligned}. \tag{31}$$

By solving the derivative of Eq. (31) w.r.t. F and letting it equal to 0, we obtain:

$$F = G_a^{-1} D_a T_a, \tag{32}$$

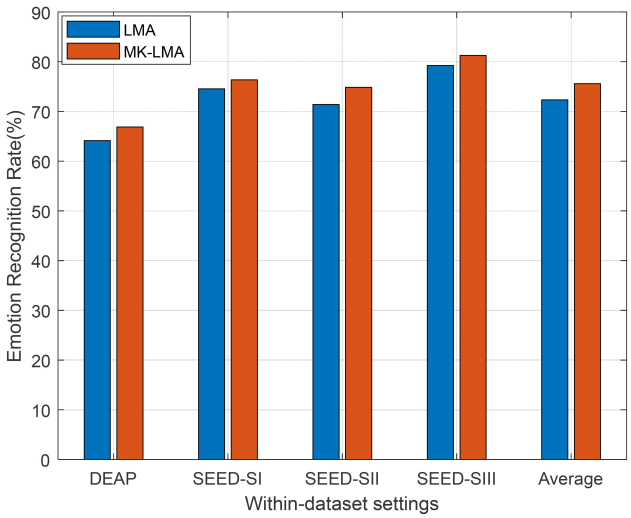

FIGURE 3 | Domain adaptation emotion recognition on within-dataset with multi-kernel learning (SI: Session I, SII: Session II, SIII: Session III).

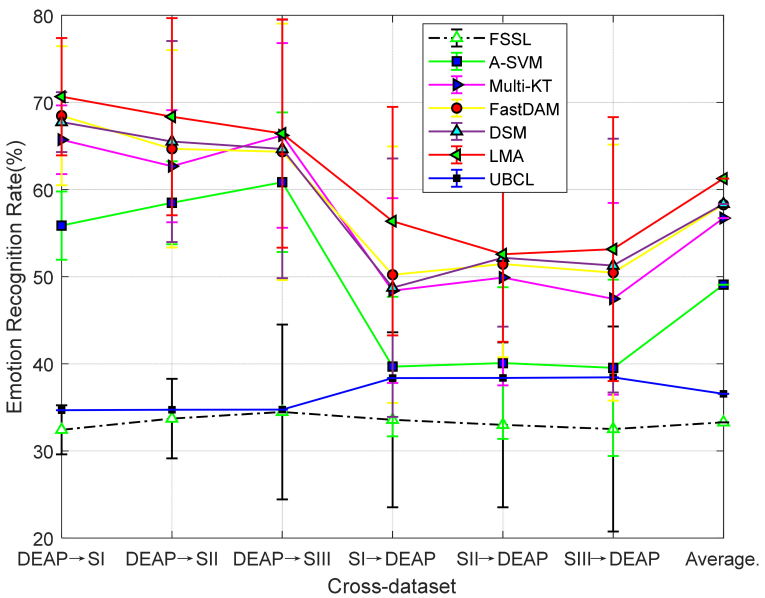

FIGURE 4 | Domain adaptation emotion recognition on cross-dataset (SI: Session I, SII: Session II, SIII: Session III).

where

$$\begin{cases} G_a = \eta_a^{q_2} S_a^{-1} \widetilde{Z}_v S_a^{-1} + \eta_a^{q_2} (S_a^{-1} - I)^T (S_a^{-1} - I) + \widetilde{\Delta} \\ D_a = \eta_a^{q_2} \left(S_a^{-1} - I\right) S_a^{-1} \widetilde{Z}_a X_t^T + \eta_a^{q_2} S_a^{-1} \widetilde{Z}_a \left(S_a^{-1} \widetilde{Z}_a - I\right) X_t^T \end{cases}, \tag{33}$$

By replacing F and F_a in Eq. (29) with those in Eqs (30) and (32), respectively, and solving the derivative of Eq. (29) with reference to T_a and equaling to zero, we then get:

$$T_a = H_a^{-1} \left(\vartheta_a^{q_1} X_a Z_a Y_a + \alpha \Theta_a P_a\right), \tag{34}$$

where:

$$\begin{aligned} H_a \; = \; & \vartheta_a^{q_1} X_a Z_a X_a^T + \alpha \left(G_a + I\right) + \eta_a^{q_2} \left((S_a^{-1} - I) G_a^{-1} D_a \right. \\ & + S_a^{-1} \widetilde{Z}_a X_t^T \left.\right)^T \left((S_a^{-1} - I) G_a^{-1} D_a + S_a^{-1} \widetilde{Z}_a X_t^T\right) \\ & + \beta D_a^T G_a^{-1} \widetilde{\Delta} G_a^{-1} D_a + \eta_a^{q_2} \left(S_a^{-1} G_v^{-1} D_v + S_a^{-1} \widetilde{Z}_a X_t^T - X_t^T\right)^T \cdot \\ & \widetilde{Z}_a \left(S_a^{-1} G_v^{-1} D_v T_v + S_a^{-1} \widetilde{Z}_a X_t^T - X_t^T\right) \end{aligned} \tag{35}$$

Let $U_a = \frac{1}{2} \left(PP^T\right)^{-\frac{1}{2}}$ and by replacing T_a in Eq. (29) with Eq. (34), and solving the derivative of Eq. (29) in reference to P_a and

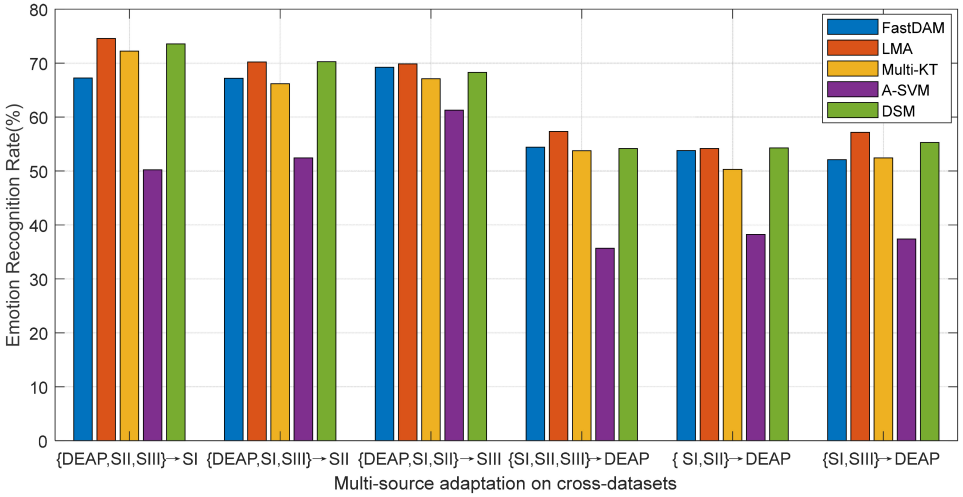

FIGURE 5 | Multi-source adaptation emotion recognition accuracy (SI: Session I, SII: Session II, SIII: Session III).

equaling to zero, we then get:

$$P_a = -\left(\Theta_a^T O_a \Theta_a + \lambda U_a\right)^{-1} M_a^T \Theta_a, \tag{36}$$

where

$$O_a = \begin{pmatrix} \vartheta_a^{q_1}\alpha^2 H_a^{-1} X_a Z_a X_a^T H_a^{-1} + \alpha^3 H_a^{-1} G_a H_a^{-1} \\ +\alpha\left(\alpha H_a^{-1} - I\right)^T \left(\alpha H_a^{-1} - I\right) \\ +\eta_a^{q_2}\alpha^2 H_a^{-1}\left((S_a^{-1} - I)G_a^{-1} D_a + S_a^{-1}\widetilde{Z}_a X_t^T\right)^T \\ \left((S_a^{-1} - I)G_a^{-1} D_a + S_a^{-1}\widetilde{Z}_a X_t^T\right) H_a^{-1} \\ +\eta_a^{q_2}\alpha^2 H_a^{-1}\left(S_a^{-1} G_v^{-1} D_v + S_a^{-1}\widetilde{Z}_a X_t^T - X_t^T\right)^T \\ \widetilde{Z}_a\left(S_a^{-1} G_v^{-1} D_v + S_a^{-1}\widetilde{Z}_a X_t^T - X_t^T\right) H_a^{-1} \\ +\beta\alpha^2 H_a^{-1} D_a^T G_a^{-1}\widetilde{\Delta} G_a^{-1} D_a H_a^{-1} \end{pmatrix}, \tag{36-1}$$

$$M_a = \begin{pmatrix} \vartheta_a^{q_1}\alpha H_a^{-1} X_a Z_a\left(\vartheta_a^q X_a^T H_a^{-1} X_a Z_a Y_a - Y\right) + \alpha^2\vartheta_a^{q_1} \\ H_a^{-1} G_a H_a^{-1} X_a Z_a Y_a + \vartheta_a^{q_1}\left(\alpha H_a^{-1} - I\right) H_a^{-1} X_a Z_a Y_a \\ +\eta_a^r\vartheta_a^{q_1}\alpha H_a^{-1}\left|(S_a^{-1} - I)G_a^{-1} D_a + S_a^{-1}\widetilde{Z}_a X_t^T\right|_F^2 H_a^{-1} \\ X_a Z_a Y_a + \beta\alpha\vartheta_a^{q_1} H_a^{-1} D_a^T G_a^{-1}\widetilde{\Delta} G_a^{-1} D_a H_a^{-1} X_a Z_a Y_a \\ +\eta_a^r\vartheta_a^{q_1}\alpha H_a^{-1}\left(S_a^{-1} G_v^{-1} D_v + S_a^{-1}\widetilde{Z}_a X_t^T - X_t^T\right)^T \widetilde{Z}_a \\ \left(S_a^{-1} G_v^{-1} D_v + S_a^{-1}\widetilde{Z}_a X_t^T - X_t^T\right) H_a^{-1} X_a Z_a Y_a \end{pmatrix}. \tag{36-2}$$

By substituting the optimal solution of the updated variables in Eqs (30), (32), (34), and (36) into Eq. (29) by mathematical calculation with the constraints $\Theta_a^T\Theta_a = I_r$, we then can get the following objective function in reference to Θ_a:

$$\Re(\Theta_a) = \min_{\Theta_a} tr\left(\Theta_a^T\left(\eta_a^{q_2} C_a - M_a\left(\Theta_a^T O_a\Theta_a + \lambda U_a\right)^{-1} M_a^T\right)\Theta_a\right), \tag{37}$$

which is equivalent to the following objective:

$$\Re(\Theta_a) = \max_{\Theta_a^T\Theta_a = I_r} tr(\Theta_a^T R_a \Theta_a), \tag{38}$$

where $R_a = M_a\left(\Theta_a^T O_a\Theta_a + \lambda U_a\right)^{-1} M_a^T - \eta_a^{q_2} C_a$. According to Li et al. (2015), Θ_a can be relaxedly obtained by the Eigen-decomposition of R_a .

Lastly, we respectively optimize ϑ_a and η_a by fixing other variables. In this situation, the objective in Eq. (29) by preserving ϑ_a changes to the following problem:

$$\min_{\vartheta_a \geq 0, \vartheta_a^T 1 = 1} \sum_{a=1}^{S} \vartheta_a^q tr(\widetilde{T}_a^T Z_a \widetilde{T}_a), \tag{39}$$

Let $g_a = tr\left(\widetilde{T}_a^T Z_a \widetilde{T}_a\right)$, the Lagrange function of Eq. (39) is

$$\Im(\vartheta_a, \varphi) = \sum_{a=1}^{S} \vartheta_a^{q_1} g_a - \varphi\left(\sum_{a=1}^{S}\vartheta_a - 1\right), \tag{40}$$

Let the derivative of $\Im(\vartheta_a, \varphi)$ with respect to ϑ_a be equivalent to 0 and we can obtain:

$$\vartheta_a = \left(\varphi / (q_1 g_a)\right)^{\frac{1}{q_1 - 1}}, \tag{41}$$

Substituting Eq. (41) into the constraint $\sum_{a=1}^{S}\vartheta_a = 1$, we obtain

$$\vartheta_a = (g_a)^{1/(1-q_1)} \Big/ \sum_{a=1}^{S}(g_a)^{1/(1-q_1)}, \tag{42}$$

With the same deduction with that of ϑ_a, we also get the following optimal solution of η_a:

$$\eta_a = (h_a)^{1/(1-q_2)} \Big/ \sum_{a=1}^{S}(h_a)^{1/(1-q_2)}, \tag{43}$$

where $h_a = tr(\Theta_a^T C_a \Theta_a) + ||F_a - F||_F^2 + tr((F_a - X_t^T T_a)^T \widetilde{Z}_a (F_a - X_t^T T_a))$.

Overall Algorithm

An overall optimization process of LMA can be outlined in the **Algorithm 1**. Following the same strategy in Zhang et al. (2019b), we employ a window-based breaking criterion to better achieve the convergence state of the algorithm. In terms of this strategy, we denote by $\hbar = 6$ the window size and compute $\varsigma = \left|MaxObj_{itr} - MinObj_{itr}\right| / MaxObj_{itr}$ in $itr-$thiteration, where $Obj_{itr} = \{Obj_{itr-\hbar+1}, ..., Obj_{itr}\}$ represents the set of historical target values in the window. While $\varsigma < \varepsilon{=}10^{-5}$, our algorithm will stop the iteration.

Algorithm 1: Multi-source adaptation learning.

Input: Source datasets $\{X_i^s\}_{i=1}^S$, Laplacian matrices $\{L_i\}_{i=1}^S$, target dataset X^t, and parameters α, β, and λ, the maximal iteration number ℓ.
Output: Converged projection matrices $\{P_i\}_{i=1}^S$, $\{\Theta_i\}_{i=1}^S$, and matrices $\{F_i\}_{i=1}^S$ and W_t.
Initialization: Set $itr = 0$, and initialize $\Theta_a = I_r$ and $\{P_a^{itr}\}_{a=1}^S$ randomly. Let $P^{itr} = [P_1^{itr}, ..., P_S^{itr}]$;
1: for $a = 1$ to S do
{
1) Compute matrix D_{itr}^a and $D_{itr}^{a(l)}$, and $\overline{K}_a^{itr}$ and $\overline{K}_{a(l)}^{itr}$ with empirical kernel mapping, thus computing
$\Lambda^a = \sum_{l=0}^c \overline{K}_a^{(l)}$, $l = 1, ..., c$ and Compute $\vartheta_a = \frac{tr\left(X_a L_a X_a^T + \Lambda^a\right)}{\sum_{a=1}^S tr\left(X_a L_a X_a^T + \Lambda^a\right)}$;
2) Initialize $T_a = \Theta_a P_a$ and $F^{itr} = \sum_{a=1}^S \vartheta_a (X^t)^T \Theta_a P_a$;
}
2: repeat
{
3) Compute W_t^{itr} by Eq. (28)
4) Compute the matrix $\widetilde{V}^{itr}$ with $(\widetilde{V})_{kk} = \frac{1}{2||(W_t^{itr})_{k,:}||_2}$;
5) set $a = 1$;
repeat
{
6) Compute the diagonal matrix Z_a^{itr}, G_a^{itr}, and $\widetilde{Z}_a^{itr}$;
7) Compute $S_a^{itr} = \widetilde{Z}_a^{itr} + I$;
8) Compute Θ_a according to Eq. (38) and then compute η_a^{itr} according to Eq. (43);
9) Compute $F_a^{itr} = (S_a^{itr})^{-1}\left(F^{itr} + \widetilde{Z}_a^{itr} X_t^T T_a^{itr}\right)$ by (30);
10) Compute F^{itr} by (32) after computing G_a^{itr} and D_a^{itr} by (33);
10) Compute T_a^{itr} by (34) with (35);
11) Compute P_a^{itr} by (36) after computing (36-1) and (36-2);
12) Compute ϑ_a^{itr} according to Eq. (42);
13) Compute the matrix
$R_a^{itr} = M_a^{itr}\left((\Theta_a^{itr})^T O_a^{itr} \Theta_a^{itr} + \lambda U_a^{itr}\right)^{-1} (M_a^{itr})^T - (\eta_a^{itr})^{q_2} C_a^{itr}$;
14) $a = a + 1$;
} **until $a > S$**
7) Update $P_a^{itr+1} = P_a^{itr}$, thus $\Theta_a^{itr+1} = \Theta_a^{itr}$ s.t. $a = 1, .., S$;
8) Update $F_a^{itr+1} = F_a^{itr}$ according to (30) s.t. $a = 1, .., S$;
9) Update ϑ_i^{itr+1} according to (42) s.t. $a = 1, .., S$;
10) Update η_i^{itr+1} according to (43) s.t. $a = 1, .., S$;
11) Update F^{itr+1} by (32), thus W_t^{itr+1} according to (28);
12) Let $itr = itr + 1$;
}**until $itr > \ell$ or $\varsigma < 10^{-5}$**
3: return $\{P_a\}_{a=1}^S$, $\{\Theta_a\}_{a=1}^S$, W_t, F and $\{F_i\}_{a=1}^S$.

In terms of the proof in Nie et al. (2010a), the convergence of the iterative procedures in **Algorithm 1** can be guaranteed by the following theorem.

Theorem 1 (Tao and Dan, 2021). The objective value in Eq. (29) would steadily decline after several iterations by **Algorithm 1**, thus finally converging to the optimum.

EXPERIMENTAL EVALUATION

In this part, we comprehensively compare the proposed method with several state of the arts on two widely used benchmark databases including SEED (Zheng and Lu, 2015) and DEAP (Koelstra et al., 2012) for EEG-based emotion recognition (Mansour et al., 2009).

Databases

According to Zhong et al. (2020) and Lan et al. (2018), there exist certain significant differences between SEED and DEAP since they can be generated by different subjects, sessions, EEG devices, experimental schemes, and emotional stimuli. Detailed information about these two datasets can be viewed in Lan et al. (2018). In the following experiments, we adopt the differential entropy (DE) (Lan et al., 2018; Zhong et al., 2020) as the data feature in emotion recognition, which has also been widely used in the preceding literatures (Shi et al., 2013; Zhang et al., 2015; Chai et al., 2016; Zheng and Lu, 2016; Chai et al., 2017; Lan et al., 2018; Zhong et al., 2020) for DA emotion recognition.

Baselines and Setting

We will systematically compare our method with such state of the arts as FSSL, an effective feature selection method without DA, FastDAM (Duan et al., 2012a), Multi-KT (Tommasi et al., 2014) with l_2-norm constraint on p, A-SVM (Yang et al., 2007), and DSM (Duan et al., 2012a). Since existing deep DA frameworks have achieved many inspiring results on emotion recognition as well as visual recognition, we also additionally present comparisons with several deep (CNN-based) DA methods with deep features: DAN (Long et al., 2015), ReverseGrad (Ganin and Lempitsky, 2015), and MultiDIAL (Carlucci et al., 2020) based on AlexNet, SDDA (Ding et al., 2018a), and CCSA (Motiian et al., 2017), a unified framework of supervised DA and generalization with deep models.

In our multi-source adaptation settings, for the baselines FSSL and A-SVM, we just equally fuse all prediction values of the base classifiers respectively obtained from each source domain[3].

In our method, LMA, there exist three vital parameters, i.e., λ, α, and β, that need to be tuned. In the community of machine learning, how to jointly search the best parameter values is still a yet unaddressed open issue. Consequently, we empirically choose these parameters using the grid search strategy also adopted in our previous work (Tao et al., 2019). Specifically, we fine-tune the values of λ, α, and β from the grid range $\{10^{-4}, 10^{-3}, ..., 10^3, 10^4\}$ in a heuristic way. Additionally, we also empirically set $q_1 = q_2 = 2$ for preventing the trivial solution in terms of the conclusion reported in Hou et al. (2017). Finally, we search the nearest neighbor number k from the set {35, 10, 15, 17}, which is also adopted in FSSL. In **Algorithm 1**, we pre-set the maximum iteration number $\tau = 100$.

Through our experiments, we adopt the RBF kernel function, i.e., $K_{i,j} = \exp(-\sigma||x_i - x_j||^2)$, for all nonlinear methods, where

[3]For each source domain, we train one SVM by using the corresponding labeled samples. Then, for each test instance x, the decision values from p SVM classifiers are converted into the probability values by using the sigmoid function (i.e., $g(t) = 1/(1 + \exp(-t))$. Finally, we average the p probability values as the final pr.

σ is equal to $1/d$. In FastDAM, we operate the same practice in Duan et al. (2012a) and set $\gamma_i = \frac{\exp(-\delta Dist(X_i^s, X))}{\sum_i \exp(-\delta Dist(X_i^s, X))}$ $(i = 1, ..., S.)$, where $\delta = 100$.

Cross-Subject Emotion Recognition

Note that different subjects even from the same dataset still have different EEG feature distributions due to the individual characteristics. We therefore practice the so-called leave-one-out cross-validation strategy conducted also in Lan et al. (2018) to evaluate the emotion recognition performance. That is, one subject remains to be the target domain, and others from the dataset are constructed as multiple sources. In this multi-source scenario, we follow the same setting as Tao and Dan (2021) to evaluate our method compared with other state of the arts on SEED and DEAP, respectively.

Performance Evaluation

We plot in **Figure 2** the recognition performance of LMA compared with the baselines on two benchmark datasets. The final obtained upper bound of chance level (UBCL) with 95% confidence interval is also recorded in **Figure 2**. It is well known that the theoretical performance (or chance level) (about 33.33 %) of the random prediction could be achieved approximately by the real chance level if the size of training data approached infinity (Lan et al., 2018). When there are finite samples, we obtain the empirical chance level by repeating the trials with the samples in question equipped with randomized class labels (Lan et al., 2018).

From **Figure 2**, we can observe that the mean performance (40.16%) of FSSL on DEAP is very close to the random prediction. While it has significantly exceeded UBCL at a 5% significance level, the relatively worse performance of FSSL still indicates the imperative importance of DA in cross-subject emotion recognition due to the substantial distribution divergence between different subjects. This importance has been witnessed by almost all baseline adaptation methods, which have yielded better performance than FSSL in all cross-subject settings. Specifically, our method, LMA, undoubtedly obtains the best recognition accuracy (about 25.14% gains over FSSL), which is closely followed by DSM. While all DA as well as our method, LMA, achieved on DEAP obvious improvement over FSSL with respect to t-test with p-value > 0.05, the mean recognition performance of these methods is yet not satisfied so far due to the complexity and difference among all subjects.

The no-adaptation method FSSL touched on SEED an average accuracy of 53.78% on three sessions from SEED, which significantly outperformed UBCL. Those multi-source adaptation methods including our method, LMA, unsurprisingly achieved more accuracy gains than the no-adaptation method on SEED. We can still observe that our method, LMA, demonstrates the best performance on SEED by upgrading the average accuracy with 75.47% w.r.t. t-test with p-value > 0.05. An interesting observation is that all methods work better on SEED than DEAP, which has also been reported in Lan et al. (2018) and Tao and Dan. (2021). The reason for this phenomenon might be that the larger distribution discrepancy between different subjects from DEAP prevented boosting performance in these methods (Mansour et al., 2009; Lan et al., 2018).

Multiple Kernel Selection

As well known, the choice of kernel is a challenging issue in the kernel learning method. Recently, multiple kernel learning (MKL) has been effectively proposed for conquering this choice issue existing in single kernel learning methods. Consequently, we also evaluate the performance boost in our method by using MKL (called as MKLMA for short) for each source domain. To this end, the first step is to construct a new space spanned by multiple kernel mapping features. We firstly denote by $\{\phi_a\}_{a=1}^{\mho}$ an empirical kernel function set, which respectively projects X_a into $\mho$ different spaces. Then, an orthogonally integrated space can be constructed by concatenating these $\mho$ spaces. We denote the mapping features in this final space by $\widetilde{\phi}(x_i) = [\phi_1(x_i)^T, \phi_2(x_i)^T, ..., \phi_{\mho}(x_i)^T]^T \in \mathbb{R}^{\mho n_a}$, where $x_i \in X_a$. Correspondingly, the kernel matrix in this final space can be easily deduced as $K_{new} = [\widetilde{K}_1; \widetilde{K}_2; ...; \widetilde{K}_{\mho}]$, where $\widetilde{K}_i$ is the ith kernel matrix from the $\mho$ feature spaces. Aiming to exploit the multiple kernel spaces, we therefore employ four kernel mapping functions including the above-used Gaussian kernel. The other additionally employed kernels are inverse square distance kernel function, Laplacian kernel function, and inverse distance kernel function, respectively, denoted as $K_{ij} = 1 \Big/ \left(1 + \sigma \left|\left|x_i - x_j\right|\right|^2\right)$, $K_{ij} = \exp\left(-\sqrt{\sigma}\left|\left|x_i - x_j\right|\right|\right)$, and $K_{ij} = 1/(1 + \sqrt{\sigma}\left|\left|x_i - x_j\right|\right|)$.

The observation from **Figure 3**, in which MKLMA significantly outperforms LMA, justifies that our LMA with MKL can further boost the recognition performance on DEAP and SEED. This also proves the importance of kernel choice in those kernel-based learning models.

Cross-Dataset Emotion Recognition

Single-Source Adaptation

In this subsection, we will demonstrate the consistent robustness of LMA by evaluating its performance in several cross-dataset settings, which is more challenging than the cross-subject adaptation due to the intrinsic difference between datasets. For the scenario of cross-dataset adaptation, we specially design several different cross-dataset strategies by splitting the training set and test set, respectively, in terms of their EEG instruments and emotional stimuli sources, thus making up six cases, i.e., *DEAP* $\rightarrow$ *SI*, *DEAP* $\rightarrow$ *SII*, *DEAP* $\rightarrow$ *SIII*, *SI* $\rightarrow$ *DEAP*, *SEED II* $\rightarrow$ *DEAP*, and *SIII* $\rightarrow$ *DEAP*, where A$\rightarrow$B denotes the adaptation from the dataset A to the dataset B, and SI, SII, and SIII are respectively denoted as the dataset of session I, session II, and session III from the database SEED.

A representative hypothesis used in DA is that the feature space of both source and target domains should be the same. Following this assumption, we employ in this part only 32 channels shared between SEED and DEAP to construct a common feature space with 160 dimensions for both domain datasets. In the first three trials, we sample 2,520 samples as the source from DEAP and 2,775 samples as the target from three different sessions (SI, SII, and SIII) in SEED. We evaluate each subject with respect to recognition accuracy in each session and

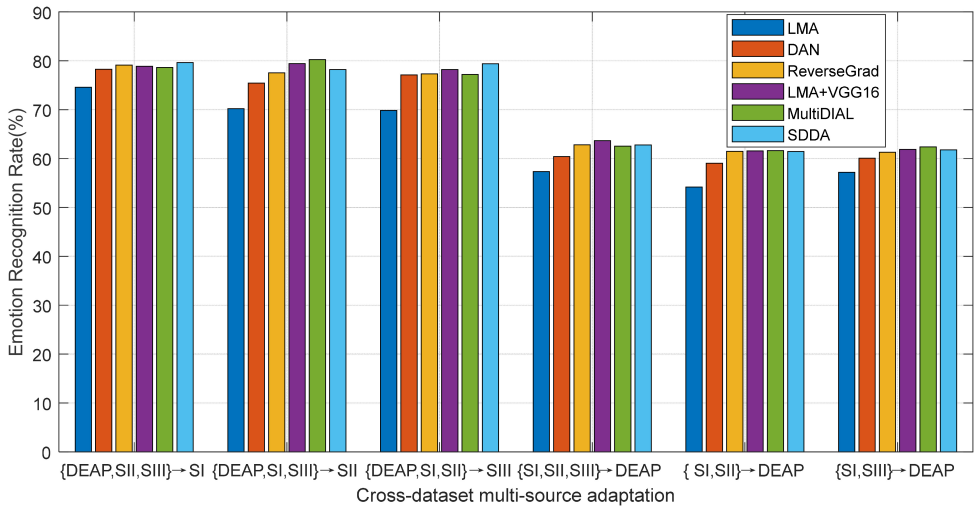

FIGURE 6 | Emotion recognition accuracies of different methods using deeply extracted features (SI: Session I, SII: Session II, SIII: Session III).

TABLE 2 | Multi-source adaptation emotion recognition accuracies of derived methods as well as LMA.

Method	{DEAP,SII,SIII} →SI	{DEAP,SI,SIII} →SII	{DEAP,SI,SII} →SIII	{SI,SII,SIII} →DEAP	{SI,SII} →DEAP	{SI,SIII} →DEAP
LMA_NF	73.52	69.10	69.58	55.43	52.01	56.22
LMA_NL	69.13	65.36	66.11	52.32	53.16	52.71
LMA_NS	72.68	68.23	68.38	54.69	51.08	54.20
LMA	**74.57**	**70.2**	**69.85**	**57.33**	**54.16**	**57.17**

Bold denotes the best recognition rates (SI: Session I, SII: Session II, SIII: Session III).

then record the final mean results over 15 subjects from SEED. In the other trials, we resample 41,625 source samples from SEED and 180 target samples from DEAP. We also record the mean recognition accuracy of each subject in DEAP over 14 subjects. For the limitation of memory, 10% of the source data (4,162 samples) is randomly sampled as the actual training samples (Shi et al., 2013; Zheng and Lu, 2015, 2016; Chai et al., 2016, 2017; Lan et al., 2018; Zhong et al., 2020).

The mean recognition results on six cross-datasets are plotted in **Figure 4**. We can observe from these results that the performance of FSSL is almost near the random guess in that it is slightly inferior to UBCL with about 95% confidence interval. Besides, as observed from the results, the mean performance of each method is slightly worse on cross-dataset than within-dataset. This confirms the larger distribution gaps between two datasets than within-dataset. The advantage of DA would be reflected in this situation since DA could potentially relieve the distribution issue in the cross-dataset applications, which can also be justified by the observation from **Figure 4**, where all DA methods outperform the no-adaptation one. While Multi-KT and FastDAM occasionally obtain the best performance in some settings, our method, LMA, still contributes the best performance in most cases.

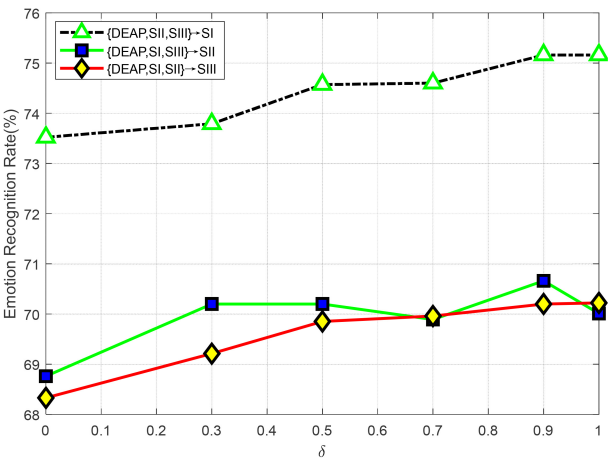

FIGURE 7 | Emotion recognition accuracies with different values of δ (SI: Session I, SII: Session II, SIII: Session III).

Multi-Source Adaptation

As reported in preceding works about DA learning, multiple source domains can improve the adaptation performance to some extent by integrating multiple prior knowledge. Nevertheless, in concrete applications, multi-source adaptation also incurs another challenge, i.e., source scalability issue, since multi-source learning could lead to the so-called "negative transfer" problem. In this scenario, how to discriminately exploit multiple sources becomes a challenge worthy to be addressed in the

multi-source adaptation learning frameworks. To this end, we will explore in this part the different reliabilities of the prior sources in the emotion recognition task (Tao et al., 2019; Tao and Dan, 2021). We evaluate the performance of all baseline DA methods with multiple prior sources on the designed cross-dataset settings. The average accuracies of all methods are plotted in **Figure 5**, where A-SVM employs the average prior model.

When there exists very large distribution discrepancy between different domain datasets, it is hard for A-SVM to eliminate the inter-domain distribution bias. Therefore, the results in **Figure 5** show that A-SVM is inferior to other multi-source adaptation methods in most settings. A-SVM even has a downgraded performance tendency with the increase of source domains in some scenarios, which indicates the existence of a "negative transfer" phenomenon in A-SVM. Another interesting observation from **Figure 5** is that all DA methods except A-SVM achieve more improvement by leveraging multiple source knowledge than by bridging only one source (i.e., cross-subject settings) when the number of source domains increase. This proves that it is beneficial to leverage multiple sources for boosting the recognition performance. Moreover, LMA and DSM conquer others by touching on the top performance, due to their designed weights for discriminately screening the optimal sources. Our method, LMA, obtains more gains over DSM in some scenarios. A possible explanation is that the shared discriminative information among source models in LMA is advantageous to multi-source adaptation learning by utilizing the optimal weight vector.

Deep Feature Adaptation

In this subsection, we will particularly evaluate our method, LMA, with deeply extracted features by comparing it with several recently proposed deep adaptation models on cross-dataset emotion recognition using the multi-source settings.

In practical tasks, our method, LMA, can be trained on the deeply transformed features of all domains, which follows the same setup with that in Zhou et al. (2018) and Zhu et al. (2017). Concretely, some pretrained deep model (e.g., VGG16) is first fine-tuned using the source domain, then the deep features can be extracted from EEG signals in both source and target domains with this CNN model, and finally the recognition model would be trained on these extracted features. In the context of our experiments, we denote our methods with the VGG16 model as LMA+VGG16. As for DAN, SDDA, MultiDIAL, and ReverseGrad, we use their released source codes to fine-tune the pre-trained models by respectively using the pre-tuned parameters in their works (Ganin and Lempitsky, 2015; Long et al., 2015). Note that these deep adaptation methods typically aim to learn domain-invariant representations. Different from the deep adaptation frameworks, our proposed method explores to learn a domain-invariant recognition model with strong generalization ability from the source domain to the target domain. Consequently, we expect that our method can further upgrade the recognition performance with the co-learning strategy on the deeply extracted features from some deep model.

We plot the mean results of all methods in **Figure 6**, from which we can observe that our deep adaptation method LMA+VGG16 significantly outperforms LMA. This indicates the advantage of deep features, which can be attributed to its robust feature representation. Furthermore, LMA+VGG16 also obtains comparable recognition performance with respect to other deep adaptation methods. This may be attributed to the classification-level constraint in LMA, where most of the source discriminative structures are expected to be preserved by the guidance of target classification. In some cases, as shown in **Figure 6**, LMA+VGG16 even achieves the top performance compared with other deep adaptation frameworks. This phenomenon shows that the proposed LMA can become an effective surrogate to the deep adaptation model by just exploiting the deep features extracted from any one of the state-of-the-art deep models.

Parameter Impact

In our method, LMA, there exist three hyper-parameters (i.e., λ, β, and α) that needed to be tuned. These hyper-parameters are mainly used to trade off different components of the LMA framework. We therefore respectively set these parameters into their extreme values to explore the importance of each component in LMA. To this end, we set $\beta = 0$ to denote LMA without target feature selection by LMA_NF, set $\alpha = \eta_a = 0$ to denote by LMA_NL the case where LMA ignores latent space representations, and set $\lambda = 0$ to denote by LMA_NS the scenario where LMA fails to consider the shared discriminative structures among multiple sources. We evaluate these derived methods on cross-dataset recognition tasks.

The results in **Table 2** clearly show that none of the three derived methods can achieve the best performance as that obtained by LMA. This further verifies the valuable contribution of each component to LMA. Specifically, LMA_NL has a significant downgraded manifestation compared with LMA, which, from the opposite side, proves that the utilization of shared latent spaces is very preferable to boosting the performance of LMA; the performance of LMA_NTF is slightly weaker than LMA, i.e., the performance of LMA would be slightly impacted by the target feature selection due to the intrinsic existence of some noise/outlier data in the target data; the inferior performance of LMA_NS proves the importance of the utilization of correlation knowledge among source models in cross-dataset emotion recognition.

Note that in section Emotion Recognition, we use the following combination function to recognize the emotion level of the given test data x_i^t:

$$j = \arg\max_j \left(y_i^t = \delta f^s(x_i^t) + (1-\delta) f^t(x_i^t)\right)_j,$$

where $\delta \in [0, 1]$ is a trade-off parameter, which is empirically set as 0.5 in the preceding trials. In this part, we will further evaluate the impact on LMA with different values of δ in multi-source adaptation scenarios. We plot the recognition accuracy w.r.t. different values of δ in **Figure 7**. From the curves shown in **Figure 7**, we can observe the following several interesting results:

(1) Theoretically, δ controls the weight of source classifiers and larger values of δ will make the source classifiers more important in LMA. An extreme case is $\delta \to 1$, where only source classifiers are guaranteed, but the target discriminative information for the test samples is discarded. In this case, all experimental results demonstrate a trend of slight downgrade. This shows the necessity of composite discrimination information by combining both source and target classifiers.
(2) Another extreme case is $\delta \to 0$. In this case, LMA will recognize the emotion state of certain test data by only using the target discriminator, which cannot leverage the prior source information with discriminating power. From **Figure 7**, we can see that all curves show an obvious upgrade in performance around $\delta = 0$, which shows the importance of multi-source discriminative models in our framework.
(3) We cannot obtain the best performance when δ values are relatively small or large, which shows the significance of exploiting the discriminative information from both source and target classifiers in our method.
(4) After $\delta > 0.5$, we can see that most curves are relatively stable across δ values, which shows that our method is not significantly sensitive to $\delta > 0.5$. Hence, we can empirically set $\delta = 0.5$ in the experiments.

CONCLUSION

To deal with the cross-subject/dataset EEG-based emotion recognition task, we proposed a robust LMA. In multiple domain-invariant latent spaces, LMA aims at transferring multi-source knowledge into target learning mainly by leveraging correlation knowledge among source models, which discriminatively screens unimportant prior evidences in sources. The comprehensive experiments performed on two public datasets verify the effectiveness of LMA in dealing with cross-subject/dataset emotion recognition. In most scenarios, our LMA (or LMA-VGG16) obtains the best results or comparable performance with respect to several representative baselines.

Since the implementation of LMA algorithm needs an iterative optimization procedure, how to improve the efficiency of LMA and seek a more efficient algorithm would be an issue worthy of further study in our future research. The unreliable and misleading pseudo labels strategy may be another potential problem in our LMA. Consequently, our successive work would explore how to seamlessly incorporate target label into the framework of LMA.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the research institution special program of Ningbo Polytechnic under Grant No. NZ21JG006, and Zhejiang Provincial Natural Science Foundation of China under Grant No. Lgg20F020013.

REFERENCES

Bao, G., Zhuang, N., Tong, L., Yan, B., Shu, J., Wang, L., et al. (2021). Two-level domain adaptation neural network for EEG-based emotion recognition. *Front. Hum. Neurosci.* 14:605246. doi: 10.3389/fnhum.2020.605246

Bruzzone, L., and Marconcini, M. (2010). Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Trans.* 32, 770–787. doi: 10.1109/TPAMI.2009.57

Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E., and Buló, S. R. (2020). MultiDIAL: domain alignment layers for (Multisource) unsupervised domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*43, 4441–4452. doi: 10.1109/TPAMI.2020.3001338

Chai, X., Wang, Q., Zhao, Y., Li, Y., Liu, D., Liu, X., et al. (2017). A fast, efficient domain adaptation technique for cross-domain electroencephalography (EEG)-based emotion recognition. *Sensors* 17:1014. doi: 10.3390/s17051014

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., and Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Comput. Biol. Med.* 79, 205–214. doi: 10.1016/j.compbiomed.2016.10.019

Chen, B., Lam, W., Tsang, I. W., and Wong, T. L. (2013). Discovering low-rank shared concept space for adapting text mining models. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1284–1297. doi: 10.1109/TPAMI.2012.243

Chu, W. S., De la Torre, F., and Cohn, J. F. (2017). Selective transfer machine for personalized facial action unit detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 529–545. doi: 10.1109/CVPR.2013.451

Ding, Z., Nasrabadi, N. M., and Fu, Y. (2018a). Semi-supervised deep domain adaptation via coupled neural networks. *IEEE Trans. Image Process.* 27, 5214–5224. doi: 10.1109/TIP.2018.2851067

Ding, Z., Nasrabadi, N. M., and Fu, Y. (2018b). Incomplete multisource transfer learning. *IEEE Trans. Neural Networks Learn. Syst.* 29, 310–323. doi: 10.1109/TNNLS.2016.2618765

Ding, Z., Sheng, L., Ming, S., and Fu, Y. (2018c). "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *Proceedings of the 15th European Conference (ECCV2018)* (Munich).

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194.

Duan, L., Tsang, I. W., and Xu, D. (2012a). Domain transfer multiple kernel learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 465–479. doi: 10.1109/TPAMI.2011.114

Duan, L., Xu, D., and Fu, C. S. (2012b). "Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1338–1345.

Duan, L., Xu, D., and Tsang, I. W. (2012c). Domain adaptation from multiple sources: a domain-dependent regularization approach. *IEEE Trans. Neural Netw. Learn. Syst.* 23, 504–518. doi: 10.1109/TNNLS.2011.2178556

Ganin, Y., and Lempitsky, V. (2015). "Unsupervised domain adaptation by back-propagation," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICMR)*, Lille, 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 1–35. doi: 10.1109/TNNLS.2020.3025954

Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. (2017). Scatter component analysis: a unified framework for domain adaptation and domain generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1414–1430. doi: 10.1109/TPAMI.2016.2599532

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). "A fast, consistent kernel two-sample test," in *Proceedings of the Conference on Neural Information Processing Systems 22*, eds Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Vancouver, BC: MIT Press), 673–681.

Hou, C., Nie, F., Tao, H., and Yi, D. (2017). Multi-view unsupervised feature selection with adaptive similarity and view weight. *IEEE Trans. Knowl. Data Eng.* 29, 1998–2011. doi: 10.1109/tkde.2017.2681670

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Comput. Intell. Magaz.* 11, 20–31. doi: 10.1109/mci.2015.2501545

Jenke, R., Peer, A., and Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 5, 327–339. doi: 10.1109/taffc.2014.2339834

Kim, M.-K., Kim, M., Oh, E., and Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from the human EEG. *Comput. Math. Methods Med.* 2013:573734. doi: 10.1155/2013/573734

Koelstra, S., Mühl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: a database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/t-affc.2011.15

Lan, Z., Sourina, O., Wang, L., Scherer, R., and Müller-Putz, G. R. (2018). Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Trans. Cogn. Dev. Syst.* 11, 85–94. doi: 10.1109/tcds.2018.2826840

Li, H., Wan, R., Wang, S., and Kot, A. C. (2021). Unsupervised domain adaptation in the wild via disentangling representation learning. *Int. J. Comput. Vis.*129, 267–283. doi: 10.1007/s11263-020-01364-5

Li, J., Qiu, S., Du, C., Wang, Y., and He, H. (2020a). Domain adaptation for EEG emotion recognition based on latent representation similarity. *IEEE Trans. Cogn. Dev. Syst.* 12, 344–353. doi: 10.1109/tcds.2019.2949306

Li, J., Qiu, S., Shen, Y.-Y., Liu, C.-L., and He, H. (2020b). Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Trans. Cyber.* 50, 3281–3293. doi: 10.1109/TCYB.2019.2904052

Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., and Hu, B. (2018a). Exploring EEG features in cross-subject emotion recognition. *Front. Neurosci.* 12:162. doi: 10.3389/fnins.2018.00162

Li, Y., Zheng, W., Cui, Z., Zhang, T., and Zong, Y. (2018b). "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)* (Palo Alto, CA).

Li, Y., Zheng, W., Cui, Z., Zong, Y., and Ge, S. (2018c). EEG emotion recognition based on graph regularized sparse linear regression. *Neural Process. Lett.* 49, 1–17. doi: 10.1109/taffc.2020.2994159

Li, Z., Liu, J., Tang, J., and Lu, H. (2015). Robust structured subspace learning for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 2085–2098.

Liu, Q., and Liu, H. (2021). Criminal psychological emotion recognition based on deep learning and EEG signals. *Neural Comput. Appl.* 33, 433–447. doi: 10.1007/s00521-020-05024-0

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning (PMLR)*, Lille, 97–105.

Long, M., Wang, J., Ding, G., Pan, S. J., and Yu, P. S. (2014). Adaptation Regularization: A General Framework for Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 26, 1076–1089. doi: 10.1109/tkde.2013.111

Lotfi, E., and Akbarzadeh-T, M.-R. (2014). Practical emotional neural networks. *Neural Netw.* 59, 61–72. doi: 10.1016/j.neunet.2014.06.012

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). "Domain Adaptation with Multiple Sources," in *Proceedings of the Conference on Neural Information Processing Systems* (Vancouver, BC: MIT Press), 1041–1048.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. (2017). "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEEInternational Conference on Computer Vision ICCV* (Venice).

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain Comput. Interfaces* 1, 66–84. doi: 10.1080/2326263x.2014.912881

Musha, T., Terasaki, Y., Haque, H. A., and Ivamitsky, G. A. (1997). Feature extraction from EEGs associated with emotions. *Artif. Life Robot.* 1, 15–19. doi: 10.1007/bf02471106

Nie, F., Huang, H., Cai, X., and Ding, C. (2010a). "Efficient and robust feature selection via Joint l2,1-norms minimization," in *Proceedings of the 22th Annual Conference Neural Information Processing Systems* (Cambridge, MA: MIT Press), 1813–1821.

Nie, F., Xu, D., Tsang, I. W., and Zhang, C. (2010b). Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction. *IEEE Trans. Image Process.* 19, 1921–1932. doi: 10.1109/TIP.2010.2044958

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* 22, 199–210. doi: 10.1109/TNN.2010.2091281

Pandey, P., and Seeja, K. (2019). "Emotional state recognition with EEG signals using subject independent approach," in *Data Science and Big Data Analytics. Lecture Notes on Data Engineering and Communications Technologies*, Vol. 16, eds D. Mishra, X. S. Yang, and A. Unal (Singapore: Springer), 117–124. doi: 10.1007/978-981-10-7641-1_10

Rosenstein, M. T., Marx, Z., and Kaelbling, L. P. (2005). "To transfer or not to transfer," in *Proceedings of the Conference on Neural Information Processing Systems*, (Cambridge, MA: MIT Press).

Shi, L. C., Jiao, Y. Y., and Lu, B. L. (2013). "Differential entropy feature for EEG- based vigilance estimation. 2013," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (Osaka), 6627–6630.

Shi, X., Guo, Z., Lai, Z., Yang, Z. Bao, and Zhang, D. (2015). A framework of joint graph embedding and sparse regression for dimensionality reduction. *IEEE Trans. Image Process.* 24, 1341–1355. doi: 10.1109/TIP.2015.2405474

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* 11, 532–541. doi: 10.3389/fnbot.2022.834952

Tao, J., Chung, F. L., and Wang, S. (2012). On minimum distribution discrepancy support vector machine for domain adaptation. *Pattern Recogn.* 45, 3962–3984. doi: 10.1016/j.patcog.2012.04.014

Tao, J., and Dan, Y. (2021). Multi-source Co-adaptation for EEG-based emotion recognition by mining correlation information. *Front. Neurosci.* 15:677106. doi: 10.3389/fnins.2021.677106

Tao, J., Wen, S., and Hu, W. (2015). L1-norm locally linear representation regularization multi-source adaptation learning. *Neural Netw.* 69, 80–98. doi: 10.1016/j.neunet.2015.01.009

Tao, J., Wen, S., and Hu, W. (2016). Multi-source adaptation learning with global and local regularization by exploiting joint kernel sparse representation. *Knowl. Based Syst.* 98, 76–94. doi: 10.1016/j.knosys.2016.01.021

Tao, J., and Xu, H. (2019). Discovering domain-invariant subspace for depression recognition by jointly exploiting appearance and dynamics feature representations. *IEEE Access* 99, 186417–186436. doi: 10.1109/access.2019.2961741

Tao, J., Zhou, D., Liu, F., and Zhu, B. (2019). . Latent multi-feature co-regression for visual recognition by discriminatively leveraging multi-source models. *Pattern Recogn.* 87, 296–316. doi: 10.1016/j.neunet.2019.02.007

Tao, J. W., Song, D., Wen, S., and Hu, W. (2017). Robust multi-source adaptation visual classification using supervised low-rank representation. *Pattern Recogn.* 61, 47–65. doi: 10.1016/j.patcog.2016.07.006

Tommasi, T., Orabona, F., and Caputo, B. (2014). Learning categories from few examples with multi-model knowledge transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 928–941. doi: 10.1109/TPAMI.2013.197

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. (2015). "Simultaneous deep transfer across domains and tasks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, 4068–4076. doi: 10.1109/ICCV.2015.463

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). "Adversarial Discriminative Domain Adaptation," in *Proceedings of the 2017 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), (Honolulu, HI), 2962–2971.

Wang, F., Zhang, W., Xu, Z., Ping, J., and Chu, H. (2021). A deep multi-source adaptation transfer network for cross-subject electroencephalogram emotion recognition. *Neural Comput. Appl.* 33, 9061–9073. doi: 10.1007/s00521-020-05670-4

Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., and Lin, S. (2006). Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* 29:40. doi: 10.1109/TPAMI.2007.12

Yang, J., Yan, R., and Hauptmann, A. G. (2007). "Cross-domain video concept detection using adaptive svms," in *Proceedings of the 15th ACM International Conference on Multimedia. ACM*, (Augsburg), 188–197.

Yang, Y., Ma, Z., Hauptmann, A. G., and Sebe, N. (2013). Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Trans. Multimedia* 15, 661–669. doi: 10.1109/tmm.2012.2237023

Zhang, K., Gong, M., and Schölkopf, B. (2015). "Multi-source domain adaptation: A causal view," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, (Menlo Park, CA), 3150–3157.

Zhang, Y., Chung, F., and Wang, S. (2020a). Clustering by transmission learning from data density to label manifold with statistical diffusion. *Knowl. Based Syst.* 193, 105330. doi: 10.1016/j.knosys.2019.105330

Zhang, Y., Chung, F. L., and Wang, S. (2019a). Takagi-sugeno-kang fuzzy systems with dynamic rule weights. *J. Intell. Fuzzy Syst.* 37, 8535–8550. doi: 10.1016/j.isatra.2017.10.012

Zhang, Y., Dong, J., Zhu, J., and Wu, C. (2019b). Common and special knowledge-driven TSK fuzzy system and its modeling and application for epileptic EEG signals recognition. *IEEE Access* 7, 127600–127614. doi: 10.1109/access.2019.2937657

Zhang, Y., Li, J., Zhou, X., Zhou, T., Zhang, M., Ren, J., et al. (2019c). A view-reduction based multi-view TSK fuzzy system and its application for textile color classification. *J. Ambient Intell. Hum. Comput.* 9, 1–11. doi: 10.1007/s12652-019-01495-9

Zhang, Y., Tian, F., Wu, H., Geng, X., Qian, D., Dong, J., et al. (2017). Brain MRI tissue classification based fuzzy clustering with competitive learning. *J. Med. Imaging Health Inform.* 7, 1654–1659. doi: 10.1006/cbmr.1996.0023

Zhang, Y., Wang, L., Wu, H., Geng, X., Yao, D., and Dong, J. (2016). A clustering method based on fast exemplar finding and its application on brain magnetic resonance images segmentation. *J. Med. Imaging Health Inform.* 6, 1337–1344. doi: 10.1166/jmihi.2016.1923

Zhang, Y., Wang, S., Xia, K., Jiang, Y., and Qian, P. (2020b). Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. *Inform. Fusion* 66, 170–183. doi: 10.1016/j.inffus.2020.09.002

Zheng, W. (2017). Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis. *IEEE Trans. Cogn. Dev. Syst.* 9, 281–290. doi: 10.1109/tcds.2016.2587290

Zheng, W. L., and Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Autonom. Ment. Dev.* 7, 162–175. doi: 10.1109/tamd.2015.2431497

Zheng, W. L., and Lu, B. L. (2016). "Personalizing EEG-based affective models with transfer learning," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, (New York, NY), 2732–2738.

Zheng, W. L., Zhang, Y. Q., Zhu, J. Y., and Lu, B. L. (2015). "Transfer components between subjects for EEG-based emotion recognition," in *Proceedings of the2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, (Xi'an), 917–922.

Zhong, P., Wang, D., and Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comput.* 99:1.

Zhou, X., Jin, K., Shang, Y., and Guo, G. (2018). visually interpretable representation learning for depression recognition from facial images. *IEEE Trans. Affect. Comput.* 11, 542–552. doi: 10.1109/TAFFC.2018.2828819

Zhu, Y., Shang, Y., Shao, Z., and Guo, G. (2017). Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Trans. Affect. Comput.* 9, 578–584. doi: 10.1109/TAFFC.2017.2650899

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.