



OPEN ACCESS

EDITED BY

Ateke Goshvarpour,
Imam Reza International University, Iran

REVIEWED BY

Man Fai Leung,
Anglia Ruskin University, United Kingdom
Hongli Chang,
Southeast University, China
Jianxing Liu,
Harbin Institute of Technology, China

*CORRESPONDENCE

Di Zhou
✉ 1008488@qq.com;
✉ 2019025@nbpt.edu.cn

[†]These authors have contributed equally to this work

RECEIVED 27 April 2023

ACCEPTED 04 October 2023

PUBLISHED 07 November 2023

CITATION

Tao J, Dan Y and Zhou D (2023) Local domain generalization with low-rank constraint for EEG-based emotion recognition. *Front. Neurosci.* 17:1213099. doi: 10.3389/fnins.2023.1213099

COPYRIGHT

© 2023 Tao, Dan and Zhou. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Local domain generalization with low-rank constraint for EEG-based emotion recognition

Jianwen Tao^{1†}, Yufang Dan^{1†} and Di Zhou^{2*†}

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

As an important branch in the field of affective computing, emotion recognition based on electroencephalography (EEG) faces a long-standing challenge due to individual diversities. To conquer this challenge, domain adaptation (DA) or domain generalization (i.e., DA without target domain in the training stage) techniques have been introduced into EEG-based emotion recognition to eliminate the distribution discrepancy between different subjects. The preceding DA or domain generalization (DG) methods mainly focus on aligning the global distribution shift between source and target domains, yet without considering the correlations between the subdomains within the source domain and the target domain of interest. Since the ignorance of the fine-grained distribution information in the source may still bind the DG expectation on EEG datasets with multimodal structures, multiple patches (or subdomains) should be reconstructed from the source domain, on which multi-classifiers could be learned collaboratively. It is expected that accurately aligning relevant subdomains by excavating multiple distribution patterns within the source domain could further boost the learning performance of DG/DA. Therefore, we propose in this work a novel DG method for EEG-based emotion recognition, i.e., Local Domain Generalization with low-rank constraint (LDG). Specifically, the source domain is firstly partitioned into multiple local domains, each of which contains only one positive sample and its positive neighbors and k_2 negative neighbors. Multiple subject-invariant classifiers on different subdomains are then co-learned in a unified framework by minimizing local regression loss with low-rank regularization for considering the shared knowledge among local domains. In the inference stage, the learned local classifiers are discriminatively selected according to their importance of adaptation. Extensive experiments are conducted on two benchmark databases (DEAP and SEED) under two cross-validation evaluation protocols, i.e., cross-subject within-dataset and cross-dataset within-session. The experimental results under the 5-fold cross-validation demonstrate the superiority of the proposed method compared with several state-of-the-art methods.

KEYWORDS

domain adaptation, subdomain generalization, emotion recognition, electroencephalogram, local learning

Introduction

In the field of affective computing research (Mühl et al., 2014), automatic emotion recognition (AER; Dolan, 2002) has received considerable attention from computer vision communities (Kim et al., 2013). Many EEG-based emotion recognition methods have been proposed so far (Musha et al., 1997; Jenke et al., 2014; Zheng, 2017; Niu et al., 2018; Pandey and Seeja, 2019; Chang et al., 2021, 2023; Zhou et al., 2022). From the viewpoint of machine learning, EEG-based AER can be modeled as a classification or regression problem (Kim et al., 2013; Zhang et al., 2017), in which state-of-the-arts for AER usually tailor their classifiers trained on multiple subjects and apply them to individual subjects. From both qualitative and empirical observations, the generalizability of AER could be attenuated partly due to the individual differences among subjects (Jayaram et al., 2016; Zheng and Lu, 2016; Lan et al., 2018). That is, the subject-independent classifier usually achieves an inferior generalization performance since emotion patterns may significantly vary from one subject to another (Pandey and Seeja, 2019). As a possible solution, subject-specific classifiers are usually impractical due to insufficient training data (Li X. et al., 2018; Zhou et al., 2022). While conspicuous progress has been made to conquer this issue by improving feature representations and learning models (Zheng and Lu, 2015; Song et al., 2018; Li et al., 2018a,b; Li Y. et al., 2019; Du et al., 2020; Zhong P. et al., 2020; Zhou et al., 2022), there still exists a long-standing challenge incurred by individual diversities in EEG-based AER. This challenge is primarily attributed to the fact that the learned classifiers should be generalized into previously unseen subjects that may obviously differ from those on which the classifiers are trained (Ghifary et al., 2017). To this end, numerous domain adaptation (DA) learning algorithms for AER have emerged by exploiting EEG features (Zheng et al., 2015; Chai et al., 2017; Li J. et al., 2019; Pandey and Seeja, 2019; Zhang et al., 2019b; Li et al., 2020; Chen et al., 2021; Dan et al., 2021; Tao et al., 2022). For instance, Pandey and Seeja (2019) and Li X. et al. (2018) successively proposed two subject invariant models for EEG-based emotion recognition; following the deep network architecture, in the researchers (Chai et al., 2016; Li H. et al., 2018; Luo et al., 2018; Li et al., 2018c, 2021; Wang et al., 2022; Zhou et al., 2022) designed several deep learning models for EEG-based emotion recognition.

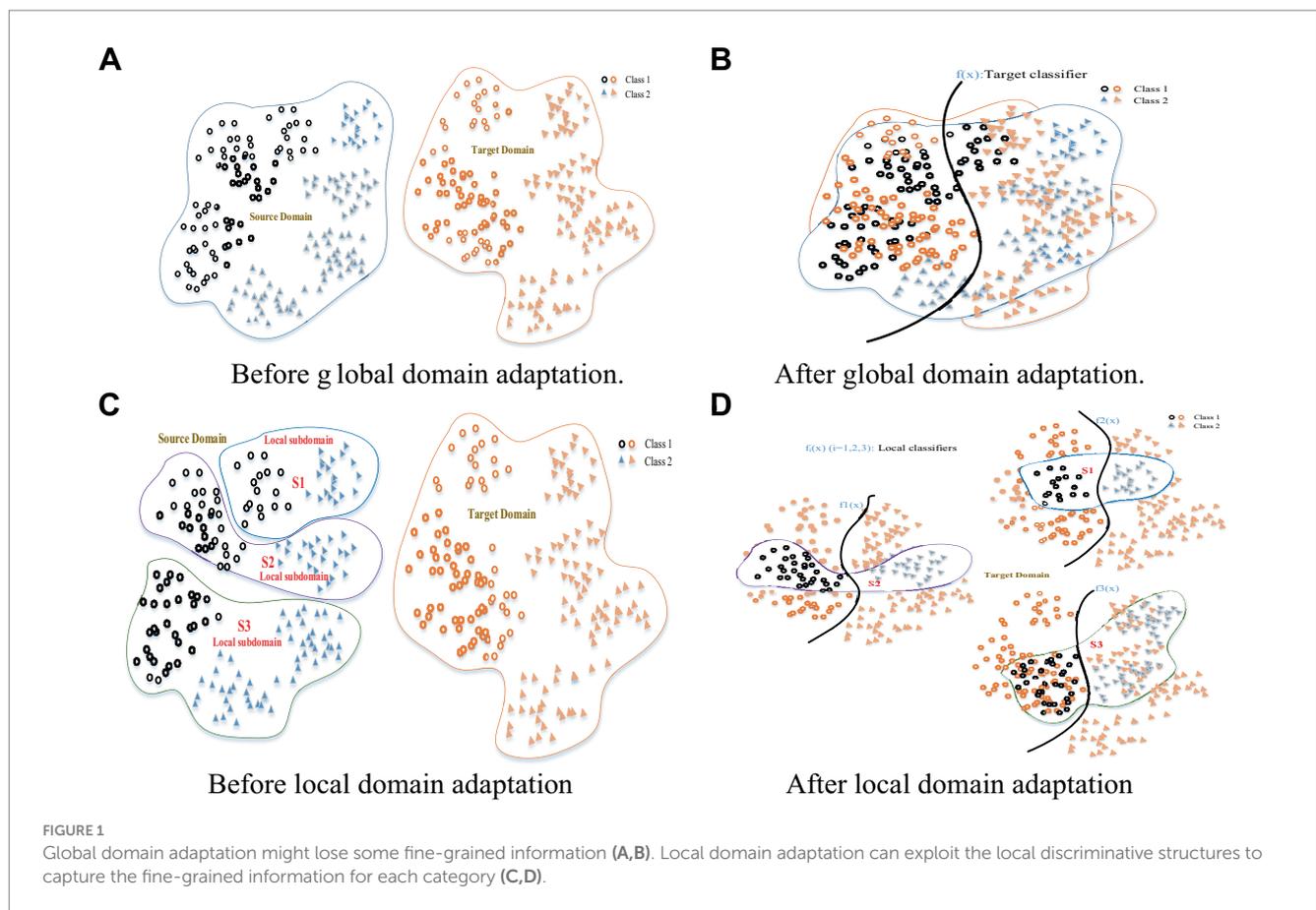
Unfortunately, in some practical AER applications, the whole target data of interest may be unavailable in the stage of training a subject-specific classifier (Wang et al., 2022). In this case, domain generalization (DG; Muandet et al., 2013), an effective variant of DA (Bruzzone and Marconcini, 2010), is proved to be a feasible solution for DA emotion recognition (Tao et al., 2022). With no need to focus on the generalization of some specific target domain, DG methodology could better acquire out-of-the-distribution effects on test samples from other previously unseen target domains (Wang et al., 2022). While DA and DG are closely related in learning scenarios, DA algorithms generally cannot be directly applicable to DG since they rely on the availability of the target domain in the stage of training. In this sense, DG is more challenging than DA as no target data can be used for fine-tuning in the training stage (Ghifary et al., 2017).

In DA/DG, one major problem is how to reduce or eliminate the distribution discrepancy between different domains (Patel et al., 2015; Wang et al., 2022). First of all, one needs to design a robust and effective criterion that can measure the domain discrepancy. Due to

its simplicity, effectiveness, and intuition, Maximum Mean Discrepancy (MMD; Gretton et al., 2009) is a commonly adopted distribution distance measure criterion. Preceding MMD-based DA methods (Pan et al., 2011; Duan et al., 2012; Tao et al., 2012, 2017, 2019; Chen et al., 2013; Long et al., 2014a; Ding et al., 2018a,b,c), however, generally focused on the global statistical distribution shift between/among different domains without considering the complementarities and diversities between two subdomains constructed with local structures within the same/different domains (Gao et al., 2015; Zhu et al., 2020). This could result in attenuated adaptation performance to some extent, since not only could all the samples from both source and target domains be confused together, but also the local discriminative structures could be trimmed without capturing the fine-grained local structures (Zhu et al., 2020). That is, while the global distribution alignment may lead to approximate zero distribution distance between different domains, a common challenge that exists in preceding global methods is that the samples from different domains are pulled too close to be accurately classified. An intuitive example is shown in Figure 1, where the source domain presents a certain multimodal structure (as shown in Figure 1A). After global domain adaptation, as shown in Figure 1B, the distributions of the two domains are approximately the same, but the data in different semantic structures are too close to be classified accurately. This is a common problem in previous global DA methods. Hence, matching the global source and target domains may not work well in this scenario.

Concerning the challenge of global domain shift, several works pay attention to semantic alignment or matching conditional distribution (Long et al., 2014a, 2017). There are other works proposed to discover multiple latent domains by decomposing the source domain (Judy et al., 2012; Gao et al., 2015). While they have presented the effectiveness of DA by exploring multiple subdomains potentially existing in the source domain, discovering multiple representative latent domains is still a non-trivial task by explicitly dividing the source samples into multiple blobs (Zhu et al., 2020). Further, to overcome the shortages that exist in the global distribution measure, numerous deep subdomain adaptation methods have focused on accurately aligning the distributions between different subdomains (Gao et al., 2015; Zhu et al., 2020). For instance, the recent work in Zhu et al. (2020) focuses on aligning the distribution of the relevant subdomains within the same category in the source and target domains. These deep learning methods, however, usually contain several updatable loss functions and converge slowly. Moreover, it is still an unexplained open problem whether the success of deep DA methods really benefits from the feature representations, fine-tuned classifiers, or effects of the adaptation regularizers (Tao et al., 2022).

Motivated by the idea of subdomain adaptation, we propose in this work a Local Domain Generalization (LDG) scheme to implicitly align the relevant local domain distributions from a single source with that of the target domain. A key improvement of LDG over previous DG/DA methods is the capability of the fine-grained alignment of a domain shift by capturing the local discriminative structures in the source domain by excavating multiple subdomains as per each positive sample with its two k-NN subsets (as shown in Figure 1C). In these local domains, multiple classifiers can be jointly trained in a unified framework by aligning them with a referenced model. Under this framework, the model discrepancies between the relevant subdomains from the source and the target domain could be measured by



considering the weights as per different distribution distances. After local domain adaptation, as shown in Figure 1D, each local domain distribution from the source domain is approximately the same as that of the target domain. Therefore, multiple local classifiers jointly learned with these local domain adaptations could be integrated and generalized into the target domain.

Specifically, we present an LDG framework for AER with EEG features with low-rank constraints. Under this framework, the source domain is firstly divided into multiple local domains, each containing only one positive sample (or exemplar; Zhang et al., 2016) and its positive and k_2 negative neighbors. Intuitively, the distribution structures of these local domains for those exemplars are expected to be relatively closer and simpler than that of the global one. In LDG, multiple subject-invariant classifiers on different local domains are co-learned in a joint framework by minimizing local regression loss. Instead of evaluating the importance of each classifier individually, LDG selects models in a collaborated mode by considering the shared knowledge among local domains by additionally imposing a nuclear-norm-based regularizer on the objective function. The learned local classifiers are discriminatively selected according to their weights in the inference stage. While the DG performance of LDG also can be boosted with most feedforward network models by exploiting the deep feature representations, it does not need iterative deep training and converges fast, thus being very efficient and effective.

Different from the existing DG methods that only focus on global distribution alignment in the source domain(s), we consider the local distribution structures of the source domain and their

relevance with the target domain to further enhance the effectiveness and generalizability of the learned adaptation model. Our algorithm can adapt as much knowledge as possible from a certain source domain, even if the EEG features between domains are partially distinct but overlapping. To the best of our knowledge, there is no prior work imposing DG with multiple local domains on solving AER problems. The main contributions of this paper are summarized as follows.

1. We propose a local domain generalization framework (LDG) for EEG-based emotion recognition by leveraging multiple structure-similar local domains from the source domain with multi-model distribution patterns. Using this framework, the capacity of MMD-based DA methods can be extended by excavating the local discriminative structures for each domain by aligning KNN-based local domain distributions.
2. We present a subdomain division strategy, i.e., splitting the source domain into multiple local domains, each of which is composed of each positive (exemplar) sample (Zhang et al., 2016; Li W. et al., 2018; Niu et al., 2018) and its k_1 positive and k_2 negative neighbors. Multiple local classifiers can be, respectively, trained on each local domain. We then formulate a new objective function by imposing a nuclear-norm-based regularizer on the model matrix in the objective function to further enhance the discriminative capability of the learned local classifiers by exploiting the intrinsic discriminative structure in the source domain.

3. An iterative optimization algorithm is presented for solving the objective of LDG that can be applied to EEG-based AER problems. The convergence of the optimization procedure can be guaranteed in terms of the proof of the proposed convergence theorem.
4. Extensive experiments are conducted on two benchmark databases (DEAP and SEED) under two cross-validation evaluation protocols (cross-subject within-dataset and cross-dataset within-session). The remarkable experimental results show that our method outperforms other state-of-the-art methods on emotion recognition tasks.

The rest of the paper is organized as follows. Section 2 reviews several related works in emotion recognition, DG, and subdomain adaptation. Section 3 introduces our LDG framework including the overall objective function, and then the optimization algorithm and its convergence analysis are successively provided in Section 4. Section 5 provides a series of experiments to evaluate the effectiveness of LDG for AER. Finally, we summarize the entire paper in Section 6.

Related work

In recent decades, increasing attention has been given to emotion recognition with brain-computer interfaces (BCI; Dolan, 2002; Kim et al., 2013; Mühl et al., 2014) in the affective computing community. A vanilla aBCI system using spontaneous EEG signals firstly extracts sufficient discriminative features from the EEG data by a certain feature extractor and then trains an optimal classifier using these features and the corresponding emotion states for AER. Therefore, a proper design of EEG-based emotion recognition models helps facilitate the data processing, benefits from discriminant feature characterization, and lightens the model performance. The latest works about affective BCI (aBCI) usually adopt machine learning algorithms on automatic emotion recognition (AER) using extracted discriminative features (Musha et al., 1997; Jenke et al., 2014; Chang et al., 2023). However, the traditional machine learning method has a major disadvantage in that the feature extraction process is usually cumbersome, and relies heavily on human experts. Then, end-to-end deep learning methods emerged as an effective way to address this disadvantage with the help of raw EEG signals and time-frequency spectrums (Han et al., 2022). More details can be found in Zhang et al. (2020c), which investigated the application of several deep learning models to the research field of EEG-based emotion recognition, including deep neural networks (DNN) (Chang et al., 2021), convolutional neural networks (CNN), long short-term memory (LSTM), and a hybrid model of CNN and LSTM (CNN-LSTM; Zhong Q. et al., 2020; Mughal et al., 2022; Xu et al., 2022).

While preceding methods have obtained remarkable achievements on EEG-based AER (Zheng, 2017; Li et al., 2018a,b; Li Y. et al., 2019; Pandey and Seeja, 2019), the performance expectation for cross-subject/dataset recognition could be lowered due to the diversities of emotional states among subjects/datasets (Jayaram et al., 2016; Zheng and Lu, 2016; Li X. et al., 2018). While subject-specific classifiers may be a possible solution for these cases, they are usually infeasible in real tasks due to insufficient training data. Moreover, even if they are

feasible in some specific scenarios, it is also an indispensable task to fine-tune the classifier to maintain a sound recognition capacity partly because the EEG signals of the same subject sometimes change (Zhou et al., 2022). Fortunately, the recently proposed domain adaptation (DA) technique (Patel et al., 2015) can be leveraged to surmount these challenges for EEG-based emotion recognition. As a well-focused research direction, the unsupervised domain adaptation (UDA) methodology has promoted a large amount of research effort devoted to generalizing the knowledge learned from one/multiple labeled source domain(s) into a different but related unlabeled target domain (Wang and Mahadevan, 2011; Gong et al., 2012; Long et al., 2014b, 2015, 2016; Ganin and Lempitsky, 2015; Ganin et al., 2016; Judy et al., 2017; Tzeng et al., 2017; Ding et al., 2018a,b,c). Over the past decade, DA-based emotion recognition methods have been a hot topic (Lan et al., 2018), almost fully covered in the literature of aBCI (Zheng et al., 2015; Chai et al., 2016, 2017; Jayaram et al., 2016; Zheng and Lu, 2016; Li H. et al., 2018; Li X. et al., 2018; Luo et al., 2018; Li et al., 2018c, 2020, 2021; Li J. et al., 2019; Chen et al., 2021; Dan et al., 2021; Tao et al., 2022; Zhou et al., 2022). Existing methods explore tackling different challenges in AER with EEG datasets by excavating a certain latent subspace shared by different domains for filling the domain distance among subjects or sessions.

In some real DA-based AER applications, the whole target data of interest may be unavailable in the stage of training (Ghifary et al., 2017). In this scenario, domain generalization (DG; Muandet et al., 2013), an effective variant of DA, has been proven to be a feasible solution for DA emotion recognition since it need not focus on the generalization of a certain specific target domain. While DA and DG are closely related in learning scenarios, DA algorithms generally are not directly applicable to DG since they rely on the availability of the target domain in the stage of training. In this sense, DG is more challenging than DA as no target data can be used for fine-tuning in the training stage. The extant works about DG can be divided into two research lines in terms of different strategies, i.e., feature-centric DG (Judy et al., 2012; Muandet et al., 2013; Ghifary et al., 2017; Motiian et al., 2017) and classifier-centric DG (Xu et al., 2014; Ghifary et al., 2015; Niu et al., 2015, 2018; Gan et al., 2016; Li W. et al., 2018). The former aims to mine domain-invariant features, while the latter uses multi-classifiers adaptation by regulating their weights. More research progress on DG can be found in the recent survey on DG (Wang et al., 2022).

As is known, a major task in vanilla UDA/DG methodology is to mitigate the domain discrepancy either by aligning the statistical moments (Pan et al., 2011; Duan et al., 2012; Tao et al., 2012; Chen et al., 2013; Long et al., 2014a,b; Xiao and Guo, 2015; Ding et al., 2018a,b,c) or by using domain adversarial learning (Ganin and Lempitsky, 2015; Ganin et al., 2016; Tzeng et al., 2017; Long et al., 2018; Pei et al., 2018) benefited from the powerful deep neural networks. Traditional DA/DG methods usually assume a global distribution shift between different domains and expect approximately the same global distribution of two domains after adaptation (Mansour et al., 2009). However, most of the preceding DA/DG methods face a common problem in that they only pay attention to matching the global statistical distribution between domains without considering the complementarities and diversities among subdomains constructed using several local structures within the same/different domains (Zhu et al., 2020). This could result in attenuated adaptation performance in part because the samples from

different domains are pulled too close to be accurately classified in those global methods. As a result, not only will all the data from the source and target domains be confused, but also the discriminative structures can be mixed up. Subdomain adaptation can to some extent conquer the shortcomings in aligning global domain discrepancy. For instance, several related works have been proposed to excavate multiple latent domains from the source domain (Judy et al., 2012). To discover multiple representative latent domains, however, is a non-trivial task done by explicitly dividing the source samples into multiple blobs. Aiming at the disadvantages of global domain adaptation, considerable works (Gao et al., 2015; Zhu et al., 2020) have explored subdomain adaptation, which focuses on aligning the local domain discrepancies. Most deep DA/DG methods belong to the deep adversarial learning methodology and converge slowly due to several loss functions. To this end, Zhu et al. (2020) recently presented a deep subdomain adaptation network (DSAN) based on the proposed local maximum mean discrepancy (LMMD), which learns a DA network by aligning the related distributions of subdomains across different domains.

It is worth noting that the discriminative structures could still be mixed up in extant subdomain adaptation schemes when the source (or target) domain presents a multimodal distribution structure (as shown in Figure 1). Different from these works on aligning global/sub-domain(s) shift(s), we propose a novel fine-grained DG method for EEG-based emotion recognition, in which multiple patches (local domains) are firstly reconstructed from the source dataset and multiple local classifiers are then learned collaboratively for effective generalization into the target domain even with multiple kinds of distribution pattern (Gao et al., 2015). Our method does not need deep training and converges fast, while its adaptation expectation can be easily boosted with deep feature representations from most feedforward network models.

Proposed framework

Preliminary notations

In the context of this paper, we, respectively, denote by small and capital letters the column vectors and matrices. The frequently used notations are summarized in Table 1. The concatenation operations of matrices along the row (horizontally) are denoted as $[A_1, A_2, \dots, A_k]$, and their concatenation along the column (vertically) are denoted as $\begin{bmatrix} A_1 \\ A_2 \\ \dots \\ A_k \end{bmatrix}$.

Definition 1 (Local domain): For a certain domain $X = \{x_i\}_{i=1}^m$ with some probability distribution P , a local domain for one positive example $x_v \in X$ is composed of its k_1 positive nearest neighbor set $N_{k_1}^+(x_v) = \{x_{v_1}, \dots, x_{v_{k_1}}\}$ and k_2 negative neighbor set $N_{k_2}^-(x_v) = \{x_{v_{k_1+1}}, \dots, x_{v_{k_1+k_2}}\}$, i.e., $X_v = \{x_v, N_{k_1}^+(x_v), N_{k_2}^-(x_v)\}$.

According to Definition 1, for any source domain $X^s = \{x_i^s\}_{i=1}^{n_s}$ with p positive samples $\{x_v^s \in \mathbb{R}^d\}_{v=1}^p$ and $n_s - p$ negative samples, one can reconstruct p local domains $X_v^s = \{x_v^s, N_{k_1}^+(x_v^s), N_{k_2}^-(x_v^s)\}$, $1 \leq v \leq p$, by finding the positive nearest neighbor set $N_{k_1}^+(x_v^s) = \{x_{v_1}^s, \dots, x_{v_{k_1}}^s\}$ and k_2 negative neighbor set for each positive sample x_v^s ($1 \leq v \leq p$).

TABLE 1 Notations and descriptions.

| Notations | Descriptions |
|--|---|
| n | Data size. |
| d | Feature dimensionality of data. |
| \mathcal{X} | Data space. |
| Γ | Label space. |
| $a = [a_1, a_2, \dots, a_d]^T \in \mathbb{R}^d$ | Feature vector. |
| $A \in \mathbb{R}^{n \times d}$ | Data matrix. |
| $A_{i,j}$ | The (i, j) entry of A . |
| A_i and A_j | The i -th row and j -th column of A . |
| A^T and a^T | The transpose of matrix A and vector a . |
| $tr(A)$ | The trace of a matrix A . |
| $\langle A_1, A_2 \rangle = tr(A_1^T A_2)$ | The inner product of two matrices A_1 and A_2 . |
| $\ a\ _p = \left(\sum_{i=1}^d a_i ^p\right)^{1/p}$ | The p -norm of a vector a . |
| $\ A\ _F = \sqrt{\sum_{i=1}^n \sum_{j=1}^d A_{i,j}^2}$ | The Frobenius norm of A . |
| I_r | Identity matrix of size $r \times r$. |
| $\mathbf{1}_d$ | d -dimensional vector of ones. |
| $\mathbf{0}_d$ | d -dimensional vector of zeroes. |

Definition 2 (Local domain adaptation, LDA): Let $\Delta = \{X_1^s, \dots, X_m^s\}$ be a set of m local domains and $X^t \notin \Delta$ be a target domain. The task of LDA is to learn an ensemble function $f_{X^t}: \mathcal{X} \rightarrow \Gamma$ by co-learning multiple classifiers $f_v(X_v^s)$ ($1 \leq v \leq m$) given Δ and X^t as the training examples by alleviating the distribution difference between source and target domains.

Definition 3 (Local domain generalization, LDG): In this scenario, the target domain is inaccessible in the training stage. Given m local domains $\Delta = \{X_1^s, \dots, X_m^s\}$, and denoted by

$X_a^s = \{x_i^a, y_i^a\}_{i=1}^{n_a}$ the samples drawn from the a -th subdomain, the task of LDG is to co-learn multiple adaptive functions $f_{X_a^s}: \mathcal{X} \rightarrow \Gamma$ only given $X_a^s, \forall a = 1, \dots, m$ as the training examples, which could be well-generalized to a certain unseen target domain.

Motivation

As is known, a major task in vanilla UDA/DG methodology is to diminish the domain discrepancy either by aligning the statistical moments (Koelstra et al., 2012; Gao et al., 2015; Li et al., 2018a, 2020) or by domain adversarial learning (Gong et al., 2012; Lan et al., 2018; Li X. et al., 2018; Ding et al., 2018a) benefited from the powerful deep neural networks (Zhu et al., 2020; Zhou et al., 2022). While extensive exploration of cross-subject/session has been conducted effectively in

the prior works by leveraging various domain adaptation tricks, one obvious shortage in these works is they usually assume a global distribution shift between different subjects and expect an approximately similar global distribution of two subjects after adaptation. In other words, these DA-based AER methods only focus on matching the global statistical distribution between subjects without considering the complementarities and diversities among local domains constructed using some intrinsic structures within the same/different subjects. This leads to attenuated adaptation performance since the real-world EEG data is usually quite diverse and the distribution of emotion data is complex. It is challenging to reduce the global distribution discrepancy between different domains.

As far as we know, limited effort, however, has been witnessed in improving DA/DG performance by leveraging local knowledge among multiple subdomains from a single source. The ignorance of the fine-grained local discriminative structures may result in unsatisfying generalization capacity in DA/DG. Exploiting the relationships among multiple local domains to match their distribution divergences could not only align the global statistical distributions but also the local discriminative patterns. In many real applications, the local structure is more important than the global structure (Ding et al., 2018a), and the local learning algorithms often outperform global learning algorithms (Ding et al., 2018b). Because of this, LDA/LDG is able to compensate for the limitation of global DA since the diversities of domain distributions intrinsically exist in real applications.

Motivated by this idea, we propose in this paper a novel domain generalization framework for EEG-based emotion recognition, i.e., Local Domain Generalization (LDG) with low-rank constraints. Under this framework, LDA is a relaxed extension of LDG, where the target domain of interest is provided during the training process. Specifically, the source domain of the auxiliary is firstly partitioned into multiple local domains, each of which contains only one positive sample (or called exemplar sample) and its k_1 positive neighbors and k_2 negative neighbors. Each local domain is expected to be relatively more similar and possess a simpler distribution structure. Then multiple subject-invariant local classifiers are co-learned on these local domains by minimizing a unified local regression loss. Instead of evaluating the importance of each classification model individually, LDG selects models in a collaborated mode for considering the shared knowledge among local domains by additionally introducing a nuclear-norm-based regularizer into the objective function. In the inference stage, the learned local classifiers are discriminatively selected and reweighted according to the distribution distance between each local domain and the target domain of interest.

In the following sections, we will present the objective formulation of our framework followed by its effective optimization algorithm.

General formulation

In LDA/LDG learning, however, there still exists two challenges worthy to be effectively addressed: (1) how to divide one source into multiple local domains and (2) how to compute the weight of each sample in its local domain. Until now, little research has been reported to address these challenges for EEG-based emotion recognition through local regression learning by decomposing the source domain into multiple local domains. To address these challenges, in this

section, we propose the general formulation of our framework underpinned by the robust local regression principle and the regularization theory. Concretely, our proposed method will possess several complementary characters, which can be combined into one unified optimization formulation so that a more effective target learning model and distribution alignment between local domains and the target domain can be jointly achieved.

For LDA of m local domains $\{X_v^s\}_{v=1}^m$ from the source domain X^s , we define the v -th ($1 \leq v \leq m$) local classifier as $f_v(w_v, X_v^s)$ corresponding to the v -th local domain, where $w_v \in \mathbb{R}^d$ is the v -th local classifier model. If we consider kernel learning and assume that there is a feature map $\phi_v: \mathcal{X} \rightarrow H_v^1$ that projects the training data from the original feature space into a certain reproducing kernel Hilbert space (RKHS; Gretton et al., 2009) H_v , then the predictor model w_v can be kernelized. We denote the kernel matrix as $(K_v)_{i,j} = \phi(x_i^v), \phi(x_j^v)$, where $x_i^v, x_j^v \in X_v^s$. We introduce the empirical kernel map as discussed in Pan et al. (2011):

$$\begin{aligned} \phi_v: \mathcal{X} &\rightarrow \mathbb{R}^d, \text{ for linear kernel mapping} \\ x &\rightarrow K_v(\cdot, x^v) \Big|_{x_1^v, x_2^v, \dots, x_n^v} = (K_v(x_1^v, x^v), \dots, K_v(x_n^v, x^v)), \\ &\text{for nonlinear kernel mapping} \end{aligned}$$

We therefore have kernelized data matrices $K_v^s = \phi_v(X_v^s)$ for nonlinear projection. For simplicity of expression, we uniformly express the data in linear and nonlinear space as follows:

$$\bar{X}_v^s = \begin{cases} X_v^s, & \text{linear} \\ K_v^s(\cdot, x), & \text{kernel} \end{cases}$$

In the sequence, we also refer to it as X_v^s (linear) and K_v^s (nonlinear) if without special denotation. We further denote by $W = [w_1; \dots; w_m]$ the concatenated local model matrix. We then endeavor to find m local adaptation models parameterized by jointly exploiting correlation information among local domains.

We first formulate our method with classical regularized empirical error (Zhang et al., 2019c), which leads to a classifier f_v based on a set of training data X_v :

$$\min \sum_{v=1}^m \text{loss}(f_v(w_v, X_v), y_v) + \Omega(f_v) \quad (1)$$

where $\Omega(f_v)$ is a regularization term that guarantees good generalization performance and $\text{loss}(\cdot, \cdot)$ is a regression loss function. Although other complex nonlinear models can be used, the linear model has the following characteristics: (1) It is fast and more suitable for practical applications and (2) The local structure of the manifold is approximately linear (Feiping Nie et al., 2010). So, we use the following linear transformation:

1 It is worthy to note that the feature mapping function ϕ_v ($1 \leq v \leq m$) with respect to each local domain can be completely different from each other.

$$f_v(w_v, X_v) = X_v^T w_v + b_v \tag{2}$$

where, $b_v \in \mathbb{R}$ is the bias term. The model vectors for all local domains should be highly correlated. So, we further get the following objective function.

$$\begin{aligned} \min_{\theta_v, w_v, b_v} & \sum_{v=1}^m \left\{ \theta_v^r \|X_v^T w_v + b_v \mathbf{1}_{k_1+k_2+1} - y_v\|_2^2 + \alpha \|w_v\|_2^2 \right\} + \beta \|W\|_* \\ \text{s.t.} & \sum_{v=1}^m \theta_v = 1, \theta_v \in [0, 1] \end{aligned} \tag{3}$$

where α, β is the regularization parameters and the coefficient θ_v is the contribution of each local model. The third term in Eq. (3) is the trace norm of $W \in \mathbb{R}^{d \times m}$, which is the convex hull of the rank of W , thus enhancing the correlation between different local weight vectors (Yang et al., 2013).

Essentially, it is expected that a bridge needs to be established between different local model vectors. Therefore, we can add a global model vector w and require each local model vector to be aligned with it (Zhang et al., 2019a). Furthermore, to avoid some noise information, we replace the real label vector y_v in Eq. (3) with the pseudo label vector $f_v \in \mathbb{R}^k$. This pseudo-label vector can be obtained by the subsequent optimization. Therefore, the objective function can be represented in the following formulation:

$$\begin{aligned} \min_{\theta_v, w_v, b_v, w} & \sum_{v=1}^n \left\{ \theta_v^r \|X_v^T w_v + b_v \mathbf{1}_k - f_v\|_2^2 + \alpha \|w_v\|_2^2 \right\} + \sum_{v=1}^n \|X_v^T w_v - X_v^T \tilde{w}\|_2^2 \\ & + \|X^T \tilde{w} + b \mathbf{1}_n - f\|_2^2 + \|f - y\|_2^2 + \beta (\|W\|_* + \|\tilde{w}\|_2^2) \\ \text{s.t.} & \sum_{v=1}^n \theta_v = 1, \theta_v \in [0, 1] \end{aligned} \tag{4}$$

where η is another regularization parameter. The reason for adding the fifth term is that the predicted results should be consistent with the actual label (Zhang et al., 2020a). We also expect that the local prediction label should be globally consistent, which is obtained by the global weight vector \tilde{w} on each local domain. In other words, the label information should be consistent with the nearby samples.

Given our objectives mentioned above, we propose the following general formulation of LDG:

$$\begin{aligned} \min_{\theta_v, w_v, f_v, b_v, \tilde{w}, \lambda_v} & \sum_{v=1}^m \left\{ \theta_v^r \|X_v^T w_v + b_v \mathbf{1}_k - f_v\|_2^2 + \alpha \|w_v\|_2^2 \right\} + \sum_{v=1}^m \|X_v^T w_v - X_v^T \tilde{w}\|_2^2 \\ & + tr \left(\tilde{w}^T \left(\sum_{v=1}^m \lambda_v X_v L_v X_v^T \right) \tilde{w} \right) \\ & + \|X^T \tilde{w} + b \mathbf{1}_n - f\|_2^2 + \|f - y\|_2^2 + \beta (\|W\|_* + \|\tilde{w}\|_2^2) + \mu \sum_{v=1}^m \lambda_v \log \lambda_v \\ \text{s.t.} & \sum_{v=1}^m \theta_v = 1, \theta_v \in [0, 1], \sum_{v=1}^m \lambda_v = 1, \lambda_v \in [0, 1] \end{aligned} \tag{5}$$

where λ_v is the contribution of different subdomains. In the above equation, the maximum entropy regularization $\lambda_v \log \lambda_v$ is added to

avoid a trivial solution. $L_v = (E_v)^{-1/2} (E_v - S_v) (E_v)^{-1/2}$ is a normalized Laplacian matrix corresponding to the v -th local domain (Yan et al., 2006), and E_v is a diagonal matrix with a diagonal element

of $(E_v)_{i,i} = \sum_j^k (S_v)_{i,j}$. The graph weight matrix S_v of X_v is defined

as follows:

$$(S_v)_{i,j} = \begin{cases} \exp\left(-\frac{x_i^v - x_j^{v2}}{\sigma^v}\right), & x_i^v \in \mathcal{N}_k(x_j^v) \text{ or } x_j^v \in \mathcal{N}_k(x_i^v) \\ 0, & \text{otherwise} \end{cases}$$

where $\mathcal{N}_k(x)$ denotes the k -NN of x .

Remark

In our objective formulation, one could adapt the knowledge obtained from multiple local domains to facilitate the target task of interest, which has been empirically demonstrated to be better than learning each local domain task independently in emotion recognition. In other words, it is expected to be beneficial to leverage the common knowledge shared by multiple local domain tasks for AER. However, most of the existing state-of-the-art algorithms uncover some optimal classifier models for the source and/or target domain independently. Moreover, in these state-of-the-art methods, joint multiple local adaptation learning has been largely unaddressed, and little or limited efforts have yet been devoted to the utilization of the correlation information among multiple local domains.

Optimization

Our objective function is non-smooth, so we propose an alternative algorithm to solve it.

Optimize b_v, w_v, f_v, f, b and \tilde{w} by fixing λ_v, θ_v .

By setting the b_v derivative to 0, we have:

$$\begin{aligned} w_v^T X_v \mathbf{1}_k + k b_v - f_v^T \mathbf{1}_k &= 0 \\ \Rightarrow b_v &= \frac{1}{k} (f_v^T \mathbf{1}_k - w_v^T X_v \mathbf{1}_k) \\ &= \frac{1}{k} (\mathbf{1}_k^T f_v - \mathbf{1}_v^T X_v^T w_v) \end{aligned} \tag{6}$$

By setting the b derivative to 0, we have:

$$\begin{aligned} \tilde{w}^T X \mathbf{1}_n + n b - f^T \mathbf{1}_n &= 0 \\ \Rightarrow b &= \frac{1}{n} (f^T \mathbf{1}_n - \tilde{w}^T X \mathbf{1}_n) \\ &= \frac{1}{n} (\mathbf{1}_n^T f - \mathbf{1}_n^T X^T \tilde{w}) \end{aligned} \tag{7}$$

Substituting Eq. (6) and Eq. (7) into Eq. (5), then setting its derivative on w_v to 0, we get the following formula:

$$w_v = Q_v^{-1} \left(\theta_v^r X_v H_k f_v + X_v X_v^T \tilde{w} \right) \tag{8}$$

where $H_k = I_k - \frac{1}{k} \mathbf{1}_k \mathbf{1}_k^T$, $Q_v = \theta_v^r X_v H_k X_v^T + X_v X_v^T + \beta V + \alpha I_d$ and $V = (W(W)^T)^{-1/2}$. By setting the derivative on \tilde{w} to 0, we get:

$$\tilde{w} = A_v^{-1} \left(X H_n f + \theta_v^r X_v X_v^T Q_v^{-1} X_v H_k f_v \right) \tag{9}$$

where $A_v = X H_n X^T - X_v X_v^T Q_v^{-1} X_v X_v^T + X_v X_v^T + \beta I_d + \lambda_v X_v L_v X_v^T$ and $H_n = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$. By setting its derivative for f_v to 0, we get:

$$f_v = B_v^{-1} \left(\theta_v^r H_k X_v^T Q_v^{-1} X_v X_v^T A_v^{-1} X H_n f \right) \tag{10}$$

where $B_v = \theta_v^r H_k - \theta_v^{2r} H_k X_v^T Q_v^{-1} X_v X_v^T A_v^{-1} X_v X_v^T Q_v^{-1} X_v$.

$$H_k - \theta_v^{2r} H_k X_v^T Q_v^{-1} X_v H_k$$

By setting its derivative for f to 0, we get:

$$f = \left(I - \theta_v^r H_n X^T A_v^{-1} X_v X_v^T Q_v^{-1} X_v H_k B_v^{-1} \theta_v^r H_k X_v^T Q_v^{-1} \right)^{-1} y \tag{11}$$

Optimize θ_v^r by fixing $b_v, w_v, f_v, \lambda_v, f, b$ and \tilde{w} .

After fixing $b_v, w_v, f_v, \lambda_v, f, b$ and \tilde{w} , the objective function in eq. (5) can be reformulated as

$$\begin{aligned} \min_{\theta_v} \sum_{v=1}^m \left\{ \theta_v^r \left\| X_v^T w_v + b_v \mathbf{1}_k - f_v \right\|_2^2 \right\} \\ \text{s.t. } \sum_{v=1}^m \theta_v = 1, \theta_v \in [0, 1], \end{aligned} \tag{12}$$

By using the Lagrange multiplier δ , we convert the above problem into a Lagrange function as follows:

$$M(\theta_v, \delta) = \sum_{v=1}^m \theta_v^2 \left(\left\| X_v^T w_v + b_v \mathbf{1}_k - f_v \right\|_2^2 \right) - \delta \left(\sum_{v=1}^m \theta_v - 1 \right) \tag{13}$$

By setting its derivative for θ_i to 0, we get:

$$\begin{aligned} \frac{\partial M}{\partial \theta_v} &= 2\theta_v \left(\left\| X_v^T w_v + b_v \mathbf{1}_k - f_v \right\|_2^2 \right) - \delta \\ \theta_v &= \frac{\delta}{2} \left(\left\| X_v^T w_v + b_v \mathbf{1}_k - f_v \right\|_2^2 \right) \end{aligned} \tag{14}$$

Since $\sum_{v=1}^m \theta_v = 1$, we obtain:

$$\theta_v = \frac{\theta_v}{\sum_{v=1}^m \theta_v} = \frac{1 / \left(\left\| X_v^T w_v + b_v \mathbf{1}_k - f_v \right\|_2^2 \right)}{\sum_{v=1}^m 1 / \left(\left\| X_v^T w_v + b_v \mathbf{1}_k - f_v \right\|_2^2 \right)} \tag{15}$$

Optimize λ_v by fixing $b_v, w_v, f_v, \theta_v, f, b$ and \tilde{w} .

When fixing $b_v, w_v, f_v, \theta_v, f, b$ and \tilde{w} , the objective function in Eq. (5) is equivalent to:

$$\begin{aligned} \min_{\lambda_v} \text{tr} \left(\tilde{w}^T \left(\sum_{v=1}^m \lambda_v X_v L_v X_v^T \right) \tilde{w} \right) + \mu \sum_{v=1}^m \lambda_v \log \lambda_v \\ \text{s.t. } \sum_{v=1}^m \lambda_v = 1, \lambda_v \in [0, 1] \end{aligned} \tag{16}$$

By using the Lagrange multiplier φ , we convert the above problem into a Lagrange function as follows:

$$\begin{aligned} L(\lambda_v, \varphi) &= \text{tr} \left(\tilde{w}^T \left(\sum_{v=1}^m \lambda_v X_v L_v X_v^T \right) \tilde{w} \right) \\ &+ \mu \sum_{v=1}^m \lambda_v \log \lambda_v - \varphi \left(\sum_{v=1}^m \lambda_v - 1 \right) \end{aligned} \tag{17}$$

By setting its derivative for λ_v to 0, we have:

$$\text{tr} \left(\tilde{w}^T X_v L_v X_v^T \tilde{w} \right) + \mu \log \lambda_v + \mu - \varphi = 0$$

We thus obtain:

$$\lambda_v = \frac{\lambda_v}{\sum_{v=1}^m \lambda_v} = \frac{\exp \left(\left(-\text{tr} \left(\tilde{w}^T X_v L_v X_v^T \tilde{w} \right) - \mu \right) / \mu \right)}{\sum_{v=1}^m \exp \left(\left(-\text{tr} \left(\tilde{w}^T X_v L_v X_v^T \tilde{w} \right) - \mu \right) / \mu \right)} \tag{18}$$

Overall algorithm and convergence analysis

According to the above objective function optimization process, we summarize the following algorithm for LDG.

Below, we will demonstrate that the alternating optimization procedure converges to the optimal solution of $\{w_v\}_{v=1}^m$ corresponding to the optimization problem (5) according to Lemma 1.

Lemma 1. For any invertible matrices M and \tilde{V} , the following inequality holds (Nie et al., 2010):

$$\frac{1}{2} \text{tr} \left(M \tilde{V}^{-\frac{1}{2}} \right) - \text{tr} \left(M^{\frac{1}{2}} \right) \geq \frac{1}{2} \text{tr} \left(\tilde{V} \tilde{V}^{-\frac{1}{2}} \right) - \text{tr} \left(\tilde{V}^{\frac{1}{2}} \right) \tag{19}$$

Next, we verify that the proposed iterative approach in Algorithm 1 can converge to the optimal solutions by the following theorem:

Theorem 1. Algorithm 1 will monotonically decrease the objective of the problem in Eq. (5) in each iteration and will converge to the optimum of the problem.

Algorithm 1: Local domain generalization and adaptation

Input: Domain training dataset $\{(X_v, y_v), v=1, \dots, m\}$; the number of nearest neighbors k_1, k_2 , and parameters α, β , and μ .

Initialization: Set $t=0$, and initialize $w_v^0, f_v^0, b_v^0, \tilde{w}^0, \theta_v^0, \lambda_v^0, b^0, f^0$ randomly, and set I_v^0 as identity matrix.

1: Construct the k -nearest neighbor graph and calculate $\{L_v\}_{v=1}^M$;

2: Compute H_k according to $H_k = I_k - \frac{1}{k} \mathbf{1}_k \mathbf{1}_k^T$;

3: Compute H_n according to $H_n = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$;

4: For each v in $\{1, \dots, m\}$

{

4.1: Let $t=0$;

4.2: repeat

{

4.2.1: Compute λ_v^t according to Eq. (18);

4.2.2: Obtain θ_v^t by Eq. (15);

4.2.3: Compute Q_v^t as $Q_v^t = (\theta_v^t)^r X_v H_k X_v^T + X_v X_v^T + \beta V^t + \alpha I_d$;

4.2.4: Update A_v^t as $A_v^t = X H_n X^T - X_v X_v^T (A_v^t)^{-1} X_v X_v^T + X_v X_v^T + \beta I_d + \lambda_v^t X_v L_v X_v^T$;

4.2.5: Update B_v^t as $B_v^t = (\theta_v^t)^r H_k - (\theta_v^t)^{2r} H_k X_v^T (Q_v^t)^{-1} X_v X_v^T (A_v^t)^{-1} X_v X_v^T (Q_v^t)^{-1} X_v H_k - (\theta_v^t)^{2r} H_k X_v^T (Q_v^t)^{-1} X_v H_k$;

4.2.6: Compute f^0 according to Eq. (11);

4.2.7: Compute f_v^t according to Eq. (10);

4.2.8: Update w as $w = A_v^{-1} \left(X H_n f^t + (\theta_v^t)^r X_v X_v^T (Q_v^t)^{-1} X_v H_k f_v^t \right)$;

4.2.9: Update w_v^t as $w_v^t = (Q_v^t)^{-1} \left((\theta_v^t)^r X_v H_k f_v^t + X_v X_v^T w \right)$;

4.2.10: Update b_v^t as $b_v^t = \frac{1}{k} \left(\mathbf{1}_k^T f_v^t - \mathbf{1}_v^T X_v^T w_v^t \right)$;

4.2.11: Update b^t as $b^t = \frac{1}{n} \left(\mathbf{1}_n^T f^t - \mathbf{1}_n^T X^T w \right)$;

4.2.12: Update V^t ;

4.2.13: Set $t = t + 1$;

} until $|\max \odot_t - \min \odot_t| / \max \odot_t < 10^{-4}$;

4.3 Next v ;

}

Output: Converged $\lambda_v, \theta_v, w_v, b_v, w, f, f_v$.

Proof. For ease of representation, we denote the updated $b_v, w_v, f_v, \theta_v, \lambda_v, b$, and \tilde{w} in each iteration as $b_v^t, w_v^t, f_v^t, \theta_v^t, \lambda_v^t, b^t$ and \tilde{w}^t , respectively. The inner loop to update in Step 2 of Algorithm 1 corresponds to the optimization of the following problem.

$$\begin{aligned} \min_{\theta_v, w_v, f_v, b_v, \tilde{w}, \lambda_v} \sum_{v=1}^{n_v} \left\{ \theta_v^r \left\| X_v^T w_v + b_v \mathbf{1}_k - f_v \right\|_2^2 + \alpha \left\| w_v \right\|_2^2 \right\} \\ + \sum_{v=1}^{n_v} \left\| X_v^T w_v - X_v^T \tilde{w} \right\|_2^2 + tr \left(\tilde{w}^T \left(\sum_{v=1}^{n_v} \lambda_v X_v L_v X_v^T \right) \tilde{w} \right) \\ + \left\| X^T \tilde{w} + b \mathbf{1}_n - f \right\|_2^2 + \left\| f - y \right\|_2^2 + \beta \left(\left\| w \right\|_* + \left\| \tilde{w} \right\|_2^2 \right) \\ + \mu \sum_{v=1}^{n_v} \lambda_v \log \lambda_v \end{aligned} \tag{20}$$

According to the definitions of the matrix V , we obtain:

$$\begin{aligned} \min_{\theta_v, w_v, f_v, b_v, \tilde{w}, \lambda_v} \sum_{v=1}^M \left\{ (\theta_v^{t+1})^r \left\| X_v^T w_v^{t+1} + b_v^{t+1} \mathbf{1}_k - f^{t+1} \right\|_2^2 + \alpha \left\| w_v^{t+1} \right\|_2^2 \right\} \\ + \sum_{v=1}^M \left\| X_v^T w_v^{t+1} - X_v^T \tilde{w}^{t+1} \right\|_2^2 + tr \left(\left(\tilde{w}^{t+1} \right)^T \left(\sum_{v=1}^M \lambda_v^{t+1} X_v L_v X_v^T \right) \tilde{w}^{t+1} \right) \\ + \left\| X^T \tilde{w}^{t+1} + b^{t+1} \mathbf{1}_n - f^{t+1} \right\|_2^2 + \left\| f^{t+1} - y \right\|_2^2 + \beta \left(\left\| w^{t+1} \right\|_* + \left\| \tilde{w}^{t+1} \right\|_2^2 \right) \\ + \mu \sum_{v=1}^M \lambda_v^{t+1} \log \lambda_v^{t+1} \\ \leq \min_{\theta_v, w_v, f_v, b_v, \tilde{w}, \lambda_v} \sum_{v=1}^M \left\{ (\theta_v^t)^r \left\| X_v^T w_v^t + b_v^t \mathbf{1}_k - f^t \right\|_2^2 + \alpha \left\| w_v^t \right\|_2^2 \right\} \\ + \sum_{v=1}^M \left\| X_v^T w_v^t - X_v^T \tilde{w}^t \right\|_2^2 + tr \left(\left(\tilde{w}^t \right)^T \left(\sum_{v=1}^M \lambda_v^t X_v L_v X_v^T \right) \tilde{w}^t \right) \\ + \left\| X^T \tilde{w}^t + b^t \mathbf{1}_n - f^t \right\|_2^2 + \left\| f^t - y \right\|_2^2 + \beta \left(\left\| w^t \right\|_* + \left\| \tilde{w}^t \right\|_2^2 \right) \\ + \mu \sum_{v=1}^M \lambda_v^t \log \lambda_v^t \end{aligned} \tag{21}$$

Eq. (21) is equivalent to the following form:

$$\begin{aligned}
 & \min_{\theta_v, w_v, f_v, \lambda_v} \sum_{v=1}^M \left\{ (\theta_v^{l+1})^r \|X_v^T w_v^{l+1} + b_v^{l+1} \mathbf{1}_k - f_v^{l+1}\|_2^2 + \alpha \|w_v^{l+1}\|_2^2 \right\} \\
 & + \sum_{v=1}^M \|X_v^T w_v^{l+1} - X_v^T \tilde{w}^{l+1}\|_2^2 + \text{tr} \left((\tilde{w}^{l+1})^T \left(\sum_{v=1}^M \lambda_v^{l+1} X_v L_v X_v^T \right) \tilde{w}^{l+1} \right) \\
 & + \|X^T \tilde{w}^{l+1} + b^{l+1} \mathbf{1}_n - f^{l+1}\|_2^2 + \|f^{l+1} - y\|_2^2 + \beta \|\tilde{w}^{l+1}\|_2^2 \\
 & + \mu \sum_{v=1}^M \lambda_v^{l+1} \log \lambda_v^{l+1} + \beta \text{tr} \left(W^{l+1} (W^{l+1})^T V^{l+1} \right) - \frac{\beta}{2} \text{tr} \left((W^{l+1} (W^{l+1})^T)^{\frac{1}{2}} \right) \\
 & + \frac{\beta}{2} \text{tr} \left((W^{l+1} (W^{l+1})^T)^{\frac{1}{2}} \right) \tag{22} \\
 & \leq \min_{\theta_v, w_v, f_v, \lambda_v} \sum_{v=1}^M \left\{ (\theta_v^l)^r \|X_v^T w_v^l + b_v^l \mathbf{1}_k - f_v^l\|_2^2 + \alpha \|w_v^l\|_2^2 \right\} \\
 & + \sum_{v=1}^M \|X_v^T w_v^l - X_v^T \tilde{w}^l\|_2^2 + \text{tr} \left((\tilde{w}^l)^T \left(\sum_{v=1}^M \lambda_v^l X_v L_v X_v^T \right) \tilde{w}^l \right) \\
 & + \|X^T \tilde{w}^l + b^l \mathbf{1}_n - f^l\|_2^2 + \|f^l - y\|_2^2 + \beta \|\tilde{w}^l\|_2^2 \\
 & + \mu \sum_{v=1}^M \lambda_v^l \log \lambda_v^l + \beta \text{tr} \left(W^l (W^l)^T V^l \right) - \frac{\beta}{2} \text{tr} \left((W^l (W^l)^T)^{\frac{1}{2}} \right) \\
 & + \frac{\beta}{2} \text{tr} \left((W^l (W^l)^T)^{\frac{1}{2}} \right)
 \end{aligned}$$

Since $V^l = \frac{1}{2} (W^l (W^l)^T)^{-\frac{1}{2}}$ and according to Lemma 1, we obtain:

$$\begin{aligned}
 & \beta \text{tr} \left(W^{l+1} (W^{l+1})^T V^{l+1} \right) - \beta \text{tr} \left((W^{l+1} (W^{l+1})^T)^{\frac{1}{2}} \right) \\
 & \geq \beta \text{tr} \left(W^l (W^l)^T V^l \right) - \beta \text{tr} \left((W^l (W^l)^T)^{\frac{1}{2}} \right) \tag{23}
 \end{aligned}$$

Subtracting (23) from (22), we have:

$$\begin{aligned}
 & \min_{\theta_v, w_v, f_v, \lambda_v} \sum_{v=1}^M \left\{ (\theta_v^{l+1})^r \|X_v^T w_v^{l+1} + b_v^{l+1} \mathbf{1}_k - f_v^{l+1}\|_2^2 + \alpha \|w_v^{l+1}\|_2^2 \right\} \\
 & + \sum_{v=1}^M \|X_v^T w_v^{l+1} - X_v^T \tilde{w}^{l+1}\|_2^2 \\
 & + \text{tr} \left((\tilde{w}^{l+1})^T \left(\sum_{v=1}^M \lambda_v^{l+1} X_v L_v X_v^T \right) \tilde{w}^{l+1} \right) \\
 & + \|X^T \tilde{w}^{l+1} + b^{l+1} \mathbf{1}_n - f^{l+1}\|_2^2 + \|f^{l+1} - y\|_2^2 + \beta \|\tilde{w}^{l+1}\|_2^2 \\
 & + \mu \sum_{v=1}^M \lambda_v^{l+1} \log \lambda_v^{l+1} + \frac{\beta}{2} \text{tr} \left((W^{l+1} (W^{l+1})^T)^{\frac{1}{2}} \right) \\
 & \leq \min_{\theta_v, w_v, f_v, \lambda_v} \sum_{v=1}^M \left\{ (\theta_v^l)^r \|X_v^T w_v^l + b_v^l \mathbf{1}_k - f_v^l\|_2^2 + \alpha \|w_v^l\|_2^2 \right\} \\
 & + \sum_{v=1}^M \|X_v^T w_v^l - X_v^T \tilde{w}^l\|_2^2 + \text{tr} \left((\tilde{w}^l)^T \left(\sum_{v=1}^M \lambda_v^l X_v L_v X_v^T \right) \tilde{w}^l \right) \\
 & + \|X^T \tilde{w}^l + b^l \mathbf{1}_n - f^l\|_2^2 + \|f^l - y\|_2^2 + \beta \|\tilde{w}^l\|_2^2 + \mu \sum_{v=1}^M \lambda_v^l \log \lambda_v^l \\
 & + \frac{\beta}{2} \text{tr} \left((W^l (W^l)^T)^{\frac{1}{2}} \right) \tag{24}
 \end{aligned}$$

The above formula is equivalent to:

$$\begin{aligned}
 & \min_{\theta_v, w_v, f_v, \lambda_v} \sum_{v=1}^M \left\{ (\theta_v^{l+1})^r \|X_v^T w_v^{l+1} + b_v^{l+1} \mathbf{1}_k - f_v^{l+1}\|_2^2 + \alpha \|w_v^{l+1}\|_2^2 \right\} \\
 & + \sum_{v=1}^M \|X_v^T w_v^{l+1} - X_v^T \tilde{w}^{l+1}\|_2^2 \\
 & + \text{tr} \left((\tilde{w}^{l+1})^T \left(\sum_{v=1}^M \lambda_v^{l+1} X_v L_v X_v^T \right) \tilde{w}^{l+1} \right) + \|X^T \tilde{w}^{l+1} + b^{l+1} \mathbf{1}_n - f^{l+1}\|_2^2 \\
 & + \|f^{l+1} - y\|_2^2 \\
 & + \beta \|\tilde{w}^{l+1}\|_2^2 + \mu \sum_{v=1}^M \lambda_v^{l+1} \log \lambda_v^{l+1} + \beta \|W^{l+1}\|_* \\
 & \leq \min_{\theta_v, w_v, f_v, \lambda_v} \sum_{v=1}^M \left\{ (\theta_v^l)^r \|X_v^T w_v^l + b_v^l \mathbf{1}_k - f_v^l\|_2^2 + \alpha \|w_v^l\|_2^2 \right\} \\
 & + \sum_{v=1}^M \|X_v^T w_v^l - X_v^T \tilde{w}^l\|_2^2 + \text{tr} \left((\tilde{w}^l)^T \left(\sum_{v=1}^M \lambda_v^l X_v L_v X_v^T \right) \tilde{w}^l \right) \\
 & + \|X^T \tilde{w}^l + b^l \mathbf{1}_n - f^l\|_2^2 + \|f^l - y\|_2^2 + \beta \|\tilde{w}^l\|_2^2 + \mu \sum_{v=1}^M \lambda_v^l \log \lambda_v^l + \beta \|W^l\|_* \tag{25}
 \end{aligned}$$

Therefore, we have proved the theorem. Because of the updating rule in Algorithm 1, the objective function shown in (5) monotonically decreases, and it is easy to see that the algorithm converges.

Target inference

After training the LDG, we get m local classifiers. In the following sections, we will separately propose ways to effectively use these learned classifiers in two cases.

1. LDG: The first is a domain generalization scenario where the target domain samples are not available during training. The other is the domain adaptation scenario with a specific target domain in which we have unlabeled data in it during the training process. In the domain generalization scenario, under the premise that we have no prior information about the target domain, we can only fuse the m local classifiers to achieve the prediction of the test sample by assigning different weights. Given a target sample x , the predictive label y can be obtained by the following formula.

$$y = \sum_{v=1}^m \theta_v^2 f_v(w_v, X_v) = \sum_{v=1}^m \theta_v^2 (x^T w_v + b_v) \tag{26}$$

2. LDA: When there is unlabeled data in the target domain, we can assign different weights to each local classifier by measuring the similarity between the target domain and each locality in the source domain to achieve a better prediction effect. In other words, when a certain local domain is closer to the target domain, we should assign a higher weight to the classifier trained on this subdomain, and vice versa.

Given a set of target domain samples $X = \{x_1, x_2, \dots, x_K\}$, where K is the number of samples in the target domain. By measuring the distance between the training sample and the target domain by the Maximum Mean Discrepancy (MMD), we get the following formula:

$$\Psi^v = \text{Dist}(X_v, X) = \frac{1}{k} \sum_{i=1}^k \phi(x_i^y) - \frac{1}{K} \sum_{j=1}^K \phi(x_j)_{H_k} \tag{27}$$

where X_v, X are the v -th local source domain and target domain datasets respectively, and $Dist(X_v, X)$ represents the distribution distance of X_v and X , and H_K denotes a regenerative kernel Hilbert space. $\phi(\cdot)$ is a Gaussian kernel nonlinear feature mapping function. Using MMD we can get the weight of each local classifier by:

$$\zeta_v = \frac{\exp(-\Psi^v)}{\sum_{v=1}^m \exp(-\Psi^v)}, v = 1, 2, \dots, m \quad (28)$$

Then we can predict the test sample x_j by the following formula:

$$y_j = \sum_v \zeta_v (x_j^T w_v + b_v) \quad (29)$$

Experimental results

In this section, we will conduct comprehensive experiments to validate the effectiveness of our method compared with several state-of-the-art ones.

Benchmark datasets

Two widely used benchmark databases, i.e., SEED (Zheng and Lu, 2015) and DEAP (Koelstra et al., 2012), are adopted for systematic experiments of EEG-based emotion recognition (Dan et al., 2021; Tao et al., 2022). More detailed descriptions of these two benchmarks can be found in Lan et al. (2018). As reported by references (Zhong P. et al., 2020; Zhong Q. et al., 2020) and (Lan et al., 2018), some obvious differences between these two benchmarks are that they may be sampled from multiple different sources such as different sessions, subjects, experimental schemes, EEG devices, and emotional stimuli, etc. Following the same practice in literature (Shi et al., 2013; Zheng et al., 2015; Chai et al., 2016, 2017; Zheng and Lu, 2016; Lan et al., 2018; Zhong P. et al., 2020; Zhong Q. et al., 2020; Tao and Dan, 2021; Tao et al., 2022) for domain adaptation emotion recognition, differential entropy (DE; Lan et al., 2018; Zhong P. et al., 2020; Zhong Q. et al., 2020) is adopted as the data feature in our experimental settings.

Baselines and protocol

Baselines

As a DG method, we compare our method with several representative domain generalization/adaptation methods, which can be summarized into the following two groups (here we only report the better models):

1. Shallow learning methods: Undo-Bias (Khosla et al., 2012), UML (Fang et al., 2013), DICA (Muandet et al., 2013), LRE-SVM (Xu et al., 2014), and SCA (Ghifary et al., 2017);
2. Deep learning methods: Deep subdomain adaptation network (DSAN; Zhu et al., 2020), Deep domain generalization with

structured low-rank constraint (DDG) (Ding et al., 2018a,b,c), deep domain confusion (DDC) (Tzeng et al., 2014), domain adversarial neural networks (DANNs) (Ganin et al., 2016), contrastive adaptation network (CAN) (Kang et al., 2022), and deep CORAL (Sun and Saenko, 2016).

Training protocol

For all datasets, we only exploit the source samples for training. We use support vector machine (SVM) as the base classifier for DICA and SCA. The tunable hyper-parameters are selected according to labels from the source domain. We adopt the Gaussian kernel with a kernel bandwidth σ computed by the median heuristic as the kernel function for the kernel-based methods. For a fair comparison, all deep learning baselines use the same architecture (ResNet101; He et al., 2016). That is, for deep domain generalization on the EEG dataset, we employed the Resnet101 architecture to extract the training features. We fine-tune all convolutional and pooling layers from pre-trained models and train the classifier layer via back-propagation. For multi-class classification of emotion recognition, we employ the ‘‘One vs. Rest’’ strategy to train our method (Zhang et al., 2020b).

Parameter setting

There are several vital parameters such as μ , α , and β that need to be determined beforehand in our objective (5) since they are employed to balance the importance of structure characterization and regularizers. Considering that parameter determination is a yet unaddressed open issue in the field of machine learning, we determine them by grid search in a heuristic way (Nie et al., 2010; Long et al., 2014b; Tao et al., 2022). Concretely, these regularization parameters are tuned from $\{10^{-4}, 10^{-3}, \dots, 10^3, 10^4\}$. Since no target labels are available for DG, it is impossible to conduct a standard cross-validation. Hence, we perform p -fold cross-validation on the labeled source subdomains, namely, calculating the averaged accuracy on each subdomain fold while exploiting the other $p - 1$ subdomain folds for training. Moreover, for constructing the nearest neighbor graph in LDG, we search the optimal neighbor number k (including k_i and k_j) in the grid $\{3, 5, 7, 9, 11, 13\}$, and then report the top-one recognition accuracy from the best parameter configuration. For the kernel learning scenarios, the Gaussian kernel, i.e., $K_{i,j} = \exp(-\sigma \|x_i - x_j\|^2)$, is used as the default kernel function, where σ is set to $1/d$ (d is the feature dimension).

Inter-subject domain generalization

Note that different subjects even from the same dataset still have different EEG feature distributions due to their characteristics. We therefore conduct the so-called leave-one-out cross-validation strategy conducted also in Lan et al. (2018) and Tao et al. (2022) to evaluate the emotion recognition performance. That is, one subject remains to be the target domain, and others from the dataset are constructed as the source domain. In this scenario, we follow the same setting as (Lan et al., 2018; Tao and Dan, 2021; Tao et al., 2022) to evaluate our method compared with other state-of-the-art methods on SEED and DEAP, respectively.

Each subject from DEAP includes 180 samples belonging to three categories, i.e., 60 samples per class. Each subject from SEED

TABLE 2 Inter-subject recognition accuracy (mean % and STD %).

| Method | | DEAP | | SEED | | | | | | | |
|---------------------------|-----------------|-------|------|-----------|------|------------|------|-------------|------|---------|------|
| | | | | Session I | | Session II | | Session III | | Average | |
| | | Mean | STD | Mean | STD | Mean | STD | Mean | STD | Mean | STD |
| Shallow methods | Undo-Bias | 60.36 | 3.41 | 69.41 | 5.44 | 65.79 | 2.24 | 72.64 | 5.10 | 69.28 | 4.26 |
| | UML | 62.18 | 4.09 | 72.57 | 6.27 | 67.58 | 1.75 | 71.17 | 3.68 | 70.44 | 3.90 |
| | DICA | 65.33 | 6.22 | 73.12 | 6.86 | 65.06 | 6.28 | 73.38 | 7.19 | 70.52 | 6.78 |
| | LRE-SVM | 68.20 | 2.12 | 77.50 | 3.29 | 70.11 | 5.44 | 77.45 | 4.53 | 75.02 | 4.42 |
| | SCA | 66.05 | 4.26 | 75.23 | 5.17 | 69.14 | 6.20 | 74.23 | 6.07 | 72.87 | 5.81 |
| | LDG | 71.51 | 3.14 | 78.92 | 5.65 | 70.88 | 5.72 | 78.93 | 5.38 | 76.24 | 5.58 |
| Deep methods | DDG | 77.68 | 3.33 | 84.92 | 6.42 | 74.29 | 7.45 | 82.33 | 8.11 | 80.51 | 7.33 |
| | DDC | 74.87 | 6.28 | 79.43 | 7.13 | 72.16 | 6.11 | 80.07 | 7.66 | 77.22 | 6.97 |
| | DANN | 75.34 | 7.11 | 82.51 | 6.49 | 73.77 | 7.59 | 83.62 | 6.51 | 79.97 | 6.86 |
| | DSAN | 78.44 | 4.15 | 84.50 | 6.18 | 74.58 | 6.33 | 84.10 | 6.12 | 81.06 | 6.21 |
| | CORAL | 74.08 | 3.58 | 80.42 | 4.20 | 71.54 | 5.49 | 81.00 | 5.00 | 77.65 | 4.90 |
| | CAN | 78.43 | 6.10 | 85.77 | 7.31 | 74.12 | 7.50 | 85.39 | 7.40 | 81.76 | 7.40 |
| | LDG + Resnet101 | 77.62 | 5.37 | 85.42 | 5.72 | 74.68 | 5.19 | 86.05 | 6.82 | 82.05 | 5.91 |
| Upp Bnd of Chn Lvl (UBCL) | | 38.85 | | 34.58 | | 34.65 | | 34.60 | | 34.61 | |

Bold denotes the best recognition rates.

contributes 2,775 samples, i.e., 925 samples per class and per session. Following the same strategy adopted by [Chai et al. \(2016\)](#), [Zheng and Lu \(2016\)](#), and [Chai et al. \(2017\)](#), we randomly sampled 1/10 of the training data (3,885 samples contributed by 14 subjects) from SEED in each experiment due to the large number of training samples. To cover the whole training dataset, we randomly extracted 10 training sets from SEED and thus conducted each experimental procedure 10 times. The final result was averaged over these 10 runs. We compared the performance of our LDG with several state-of-the-art DG approaches. The mean recognition accuracies of LDG compared with the baselines on two benchmark datasets are recorded in [Table 2](#).

As is known, when the size of training data increases to infinity, the theoretical performance (about 33.33%) of the random prediction can be approximately approached by the real chance level ([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](#)). For comparison, we also provided the upper bound of chance levels (UBCL) with a 95% confidence interval in this table.

Comparison with shallow methods

As observed from [Table 2](#), the mean performance of all methods on two datasets has significantly exceeded UBCL at a 5% significance level. This indicates the imperative importance of inter-subject domain generalization due to the intrinsic existence of distribution divergence among different subjects. Compared with other shallow learning methods, our method LDG undoubtedly obtains the best mean performance ($75.06\% \pm 4.97$) in all cases, which is followed by LRE-SVM ($73.32\% \pm 3.85$). This may be attributed to the subdomain learning technologies in LDG and LRE-SVM. Our method LDG unsurprisingly achieved more performance gains than LRE-SVM on

both DEAP and SEED. The multi-source generalization method SCA and DICA are found to be more effective than Undo-bias and UML. The experimental results in [Table 2](#) show that while the relative improvement over vanilla DA/DG methods is significant (t -test, value of $p > 0.05$), the absolute accuracy is still rather low, which suggests that there still exists adverse impact incurred by substantial distribution discrepancies between different subjects.

An interesting result that can be observed from [Table 2](#) is that all methods demonstrate better performance on SEED than on DEAP. The same observation has also been reported in [Lan et al. \(2018\)](#) and [Tao and Dan \(2021\)](#). A possible explanation for this result might be that there exist large discrepancies among different subjects, and the samples are distributed more “orderly” in their original feature space on SEED than that on DEAP ([Mansour et al., 2009](#)), thus leading to better alignment on SEED in some kernel space. That is, larger discrepancies among different subjects from DEAP could degrade the recognition accuracy to some extent ([Mansour et al., 2009](#); [Lan et al., 2018](#)).

Comparison with deep methods

Following the same settings in [Donahue et al. \(2014\)](#) and [Zhou et al. \(2022\)](#), our method LDG also can be easily trained with the deeply extracted features via the classical deep models such as VGG ([Simonyan and Zisserman, 2014](#)) and ResNet ([He et al., 2016](#)). Specifically, one can fine-tune some pre-trained deep models (e.g., Resnet101; [He et al., 2016](#)) through the source domain, and then extract the deep features from EEG signals. Finally, the recognition model can be learned using these deep representations.

In this part, we will particularly evaluate our method LDG with deeply extracted features by comparing it with several recently proposed deep adaptation models. We additionally denote our

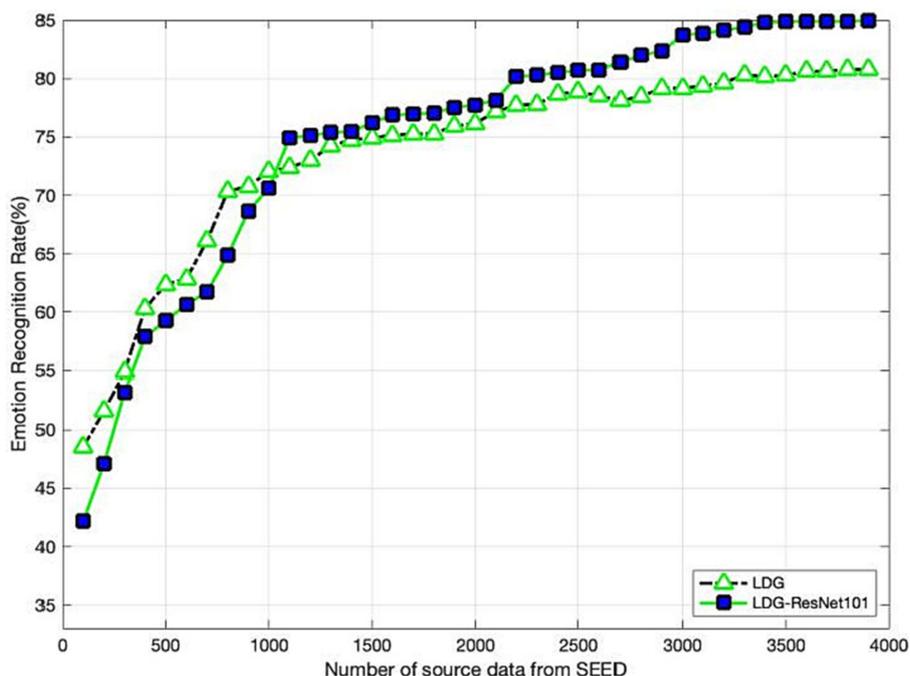


FIGURE 2 Recognition accuracy with varying sizes of source samples on SEED.

method with deep features as LDG+ResNet101. As for other deep benchmarks, their opened source codes are directly borrowed to fine-tune the pre-trained models adopted in their works, respectively. Different from these deep adaptation models, which typically pursue gaining certain domain-invariant representations, our proposed method explores optimizing a domain-invariant recognition model with strong generalization ability from the single source domain to the unseen target. We expect our method leveraging the deeply extracted features can further upgrade the recognition performance with the proposed subdomain generalization strategy.

As shown in Table 2, all of the deep methods obviously outperform the shallow ones. This indicates the advantage of deep features due to their more discriminative representations. As expected, LDG+ResNet101 also obtains the best or comparable recognition performance compared with other deep adaptation methods, followed by CAN and DSAN. This may be partly attributed to the classification-level modeling in our LDG, where most of the local discriminative structures have been preserved by the guidance of subdomain construction. In some scenarios, shown in Table 2, LDG+ResNet101 even achieves the top-one accuracy, which verifies that the proposed LDG can become an effective surrogate to the deep adaptation model by exploiting the deeply extracted features from some pre-trained deep models.

Sample size impact

Figure 2 clearly reports the impact on the performance with different sizes of source on SEED, where the source size varies from 100 to 3,800. We can observe that our methods LDG and LDG+ResNet101 manifest the same trends of upgrade in the curves.

As expected, larger source samples are beneficial to improve the recognition performance of our methods. It is worth noting that the performance of LDG can be smoothly and steadily improved with the increase of the source size, while LDG+ResNet101 can achieve significant performance when the source samples are relatively large, e.g., larger than 1,100. When the number of source samples increases to 3,500, LDG and LDG-ReNet101 asymptotically approach their equilibrium states.

Multiple kernel selection

As an open problem, how to choose an effective kernel is a challenge for learning a kernel machine. Fortunately, the previously proposed multiple kernel learning (MKL) trick can be adapted to overcome this confusion. In the sequence, we further evaluate the performance improvement in our method via introducing MKL (denoted by LDG-mkl for short) for each subdomain, in which a new feature space spanned by multiple kernel projections will be first constructed. Specifically, given an empirical kernel function set $\{\phi_a\}_{a=1}^{\mathfrak{U}}$, we, respectively, project them into \mathfrak{U} different spaces, and then construct an orthogonally integrated feature space by horizontally concatenating these spaces. In this concatenated space, the projected features can be denoted by

$$\tilde{\phi}(x_i) = [\phi_1(x_i)^T, \phi_2(x_i)^T, \dots, \phi_{\mathfrak{U}}(x_i)^T]^T \in \mathbb{R}^{\mathfrak{U}n_s}$$
, where $x_i \in X_a$, and then the kernel matrix can be easily deduced as $K_{new} = [\tilde{K}_1; \tilde{K}_2; \dots; \tilde{K}_{\mathfrak{U}}]$, where \tilde{K}_i is the i -th kernel matrix from the \mathfrak{U} feature spaces. Following the same strategy in Long et al. (2015), besides the above-used Gaussian kernel, we additionally introduce another three kernel functions including inverse

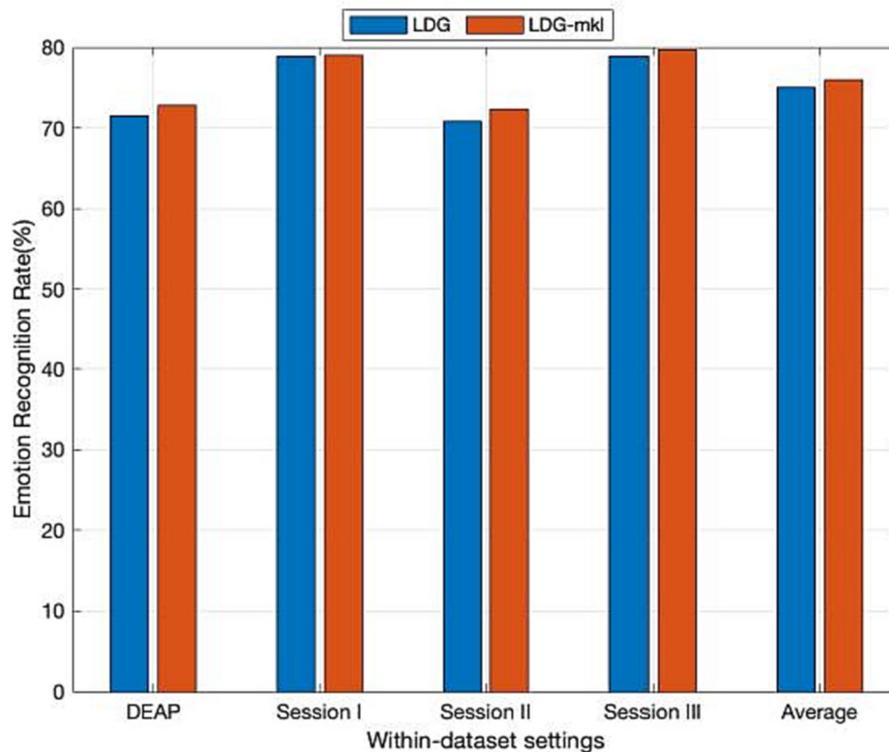


FIGURE 3 Domain adaptation emotion recognition on within-dataset with multi-kernel learning (SI: Session I, SII: Session II, SIII: Session III).

square distance kernel function, Laplacian kernel function, and inverse distance kernel function, which are, respectively, denoted as $K_{ij} = 1 / (1 + \sigma \|x_i - x_j\|^2)$, $K_{ij} = \exp(-\sqrt{\sigma} \|x_i - x_j\|)$, and $K_{ij} = 1 / (1 + \sqrt{\sigma} \|x_i - x_j\|)$. The observed mean experimental results from Figure 3 prove that LDG-mkl can boost the performance of LDG with a single kernel. This also verifies that the performance improvement in the kernel machines can be attributed to the diversities of multiple kernel functions.

Cross-dataset domain generalization

In this subsection, we further evaluate the broad and consistent generalization capacity of our LDG method on cross-dataset emotion recognition. Intuitively speaking, cross-data generalization must be more challenging than cross-subject generalization due to the significant difference between datasets.

Following the same settings in Tao and Dan (2021) and Tao et al. (2022), for robust cross-dataset generalization, the 32 shared channels by SEED and DEAP are employed to support a common feature space of 160 dimensions. In this case, several cross-dataset generalization settings can be made up, i.e., $DEAP \rightarrow SI$, $DEAP \rightarrow SII$, $DEAP \rightarrow SIII$, $SI \rightarrow DEAP$, $SEED II \rightarrow DEAP$, and $SIII \rightarrow DEAP$, where “ $x \rightarrow y$ ” means domain generalization from the dataset x into the dataset y , and SI, SII, and SIII are, respectively, denoted as the Session I, Session II, and Session III from SEED. When DEAP is

regarded as the source dataset, 2,520 data are sampled from DEAP and 2,775 data taken as the target datasets are, respectively, sampled from three different sessions (SI, SII, and SIII) of SEED. When each session of SEED is taken as the source dataset, we resample 41,625 data from it as a training set and 180 samples from DEAP regarded as the target dataset. We report the mean generalization results on six cross-dataset in Table 3.

It can be seen from the experimental results in Table 3 that the average performance of each method on the cross-dataset is slightly worse than that in the within-dataset. This confirms that the distribution difference between the two datasets is greater than that between the two subjects. The superiority of subdomain generalization will be reflected in this scenario because subdomains can potentially alleviate the distribution diversity in cross-datasets when the target dataset is unavailable in the phase of training. This can also be proved by the observation from Table 3, where our method LDG outperforms other shallow methods in almost all cases. Although SCA occasionally achieves the best performance in one setting ($SI \rightarrow DEAP$), our LDG method still achieves the top-one performance in other cases. In deep learning scenarios, all methods still undoubtedly outperform their shallow counterparts, which can be attributed to the advantage of deep feature representations. It is worth noting that our deep method LDG+Resnet101 also obtains the best or comparable recognition performance compared with other deep adaptation models. This once again proves the importance of the classification-level constraint in LDG.

Regarding the previously reported results in Yang et al. (2007), Tommasi et al. (2014), Tao et al. (2017, 2019, 2022), Ding et al.

TABLE 3 Domain adaptation emotion recognition on cross-dataset.

| Methods | | DEAP → SI | DEAP → SII | DEAP → SIII | SI → DEAP | SII → DEAP | SIII → DEAP |
|-----------------|----------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Shallow methods | Undo-Bias | 44.35 | 49.72 | 43.71 | 42.57 | 41.99 | 42.51 |
| | UML | 45.63 | 49.98 | 49.67 | 44.91 | 42.48 | 43.53 |
| | DICA | 47.35 | 52.68 | 52.11 | 43.34 | 44.90 | 42.46 |
| | LRE-SVM | 50.48 | 56.46 | 57.11 | 46.34 | 47.20 | 47.46 |
| | SCA | 48.89 | 54.35 | 54.65 | 46.73 | 45.36 | 44.67 |
| | LDG | 52.62 | 57.66 | 57.83 | 45.60 | 47.89 | 49.76 |
| Deep methods | DDG | 62.40 | 64.92 | 73.92 | 64.29 | 54.29 | 53.33 |
| | DDC | 60.89 | 62.43 | 69.43 | 62.16 | 52.16 | 50.07 |
| | DANN | 61.08 | 62.51 | 72.51 | 63.77 | 53.77 | 52.62 |
| | DSAN | 63.28 | 64.50 | 74.50 | 64.58 | 55.58 | 54.10 |
| | CORAL | 60.15 | 60.42 | 70.42 | 61.54 | 52.54 | 51.00 |
| | CAN | 64.22 | 65.77 | 75.77 | 66.12 | 57.12 | 55.39 |
| | LDG+ Resnet101 | 62.62 | 65.81 | 74.42 | 66.86 | 55.68 | 56.18 |

Bold denotes the best recognition rates.

TABLE 4 Multi-dataset generalization (SI: Session I, SII: Session II, SIII: Session III).

| Methods | {DEAP,SII,SIII} → SI | {DEAP,SI,SIII} → SII | {DEAP,SI,SII} → SIII | {SI,SII,SIII} → DEAP | {SI,SII} → DEAP | {SI,SIII} → DEAP |
|-----------|----------------------|----------------------|----------------------|----------------------|-----------------|------------------|
| Undo-Bias | 46.16 | 51.32 | 45.11 | 43.84 | 40.68 | 41.76 |
| UML | 44.06 | 50.10 | 51.21 | 46.01 | 42.90 | 43.85 |
| DICA | 49.28 | 52.94 | 54.06 | 46.62 | 45.39 | 44.63 |
| LRE-SVM | 47.17 | 57.30 | 59.30 | 50.77 | 46.50 | 49.06 |
| SCA | 52.33 | 57.66 | 57.29 | 48.68 | 47.72 | 48.93 |
| LDG | 52.76 | 57.66 | 61.43 | 49.34 | 47.03 | 49.48 |

Bold denotes the best recognition rates.

(2018a,b,c), and Tao and Dan (2021), multi-dataset adaptation can be improved by ensemble multiple auxiliary datasets. Please note that the scalability challenge could be incurred in case of multi-dataset generalization in that multi-dataset learning could bring the so-called “negative transfer” problem (Rosenstein et al., 2005), an open issue that exists in vanilla multi-source DA (Li J. et al., 2019; Chen et al., 2021; Tao and Dan, 2021). Therefore, we particularly evaluate the scalability of the proposed method by leveraging multiple source datasets for cross-dataset generalization. We report the average performance in Table 4 on all source datasets for the single-source methods including our LDG as well as LRE-SVM, Undo-Bias, and UML.

As shown in Table 4, due to the significant distribution differences among different source datasets, it is difficult for the single-source methods to generalize to unseen target domains in multi-source datasets. Therefore, the results in Table 4 indicate that these methods are generally inferior to other multi-source fusion methods. In some scenarios, they even exhibit a performance degradation trend as the number of source domains increases, indicating the “negative transfer” phenomenon (Rosenstein et al., 2005). Another interesting observation in Table 4 is that all multi-source methods achieve slight improvements by utilizing multiple sources as opposed to bridging only a single source (i.e., cross-dataset settings) as the

number of source domains increases. This demonstrates the benefits of using multiple sources to enhance identification performance. In addition, SCA and DICA outperform other methods by conquering top-level performance as their designed weights are used to differentially screen the best sources. In some cases, our LDG method achieves more benefits than SCA. One possible explanation is that the discriminative information shared among sub-domain models in LDG is advantageous for multi-source generalization.

Convergence

Since our LDG is an iterative algorithm, it is crucial to evaluate its convergence. We evaluate the convergence of the LDG algorithm by conducting several experiments for emotion recognition in three settings such as cross-subject within DEAP, DEAP→SI, and {SI, SII, SIII} → DEAP. We plotted the mean experimental results in Figure 4. The curves in this figure show that the proposed algorithm has a certain asymptotic convergence. The objective values of LDG usually converge in less than 30 iterations. We also observed a similar phenomenon from other recognition tasks with different cross-subject/cross-dataset settings.

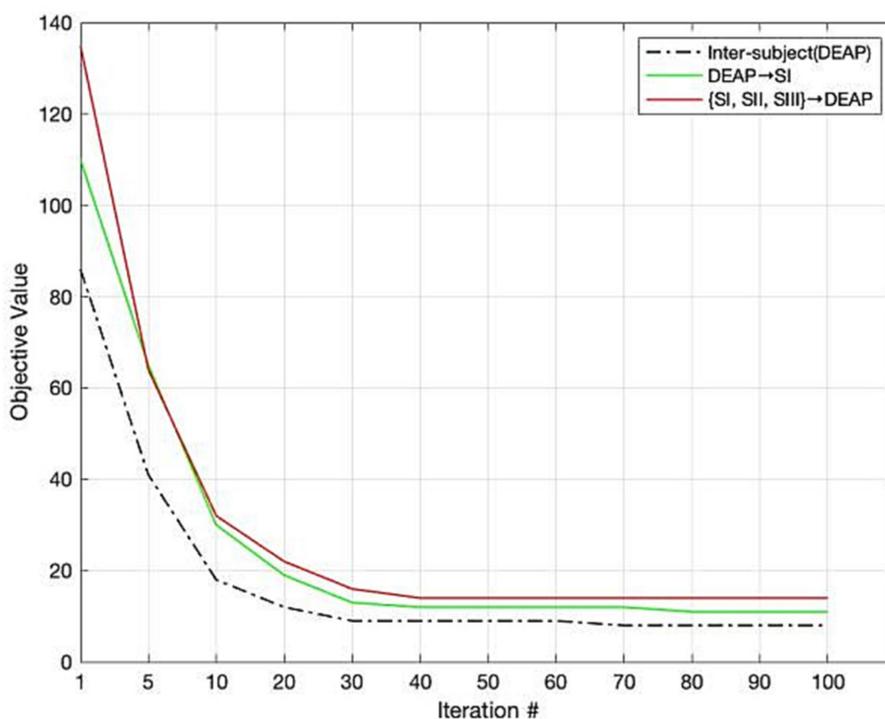


FIGURE 4
Convergence of LDG.

Conclusion

To deal with cross-subject/dataset EEG-based emotion recognition tasks, we proposed a local domain generalization (LDG) framework. In multiple subdomain spaces, LDG aims at transferring local knowledge into target learning mainly by leveraging correlation knowledge among subdomain models via low-rank constraint on the local models, which discriminatively screens unimportant prior evidence in subdomains. The comprehensive experiments performed on two public datasets verify the effectiveness of LDG in dealing with cross-subject/dataset emotion recognition. In most scenarios, our LDG and LDG+Resnet101 obtain the best results or comparable performance concerning several representative baselines.

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

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: in our research, the datasets DEAP and SEED can be, respectively, accessed from <http://epileptologie-bonn.de/cms/upload/workgroup/lehnertz/eegdata.html> and <http://bcmi.sjtu.edu.cn/~seed>.

Author contributions

DZ extensively conducted all experiments in the paper. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Ningbo Natural Science Foundation project (No. 2022J180).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Bruzzone, L., and Marconcini, M. (2010). Domain adaptation problems: a DASVM classification technique and a circular validation strategy. *IEEE Trans. PAMI* 32, 770–787. doi: 10.1109/TPAMI.2009.57
- 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
- Chang, H., Zong, Y., Zheng, W., Tang, C., Zhu, J., and Li, X. (2021). Depression assessment method: an EEG emotion recognition framework based on spatiotemporal neural network. *Front. Psychiatry* 12:837149. doi: 10.3389/fpsy.2021.837149
- Chang, H., Zong, Y., Zheng, W., Xiao, Y., Wang, X., Zhu, J., et al. (2023). EEG-based major depressive disorder recognition by selecting discriminative features via stochastic search. *J. Neural Eng.* 20:026021. doi: 10.1088/1741-2552/acbe20
- Chen, H., Jin, M., Li, Z., Fan, C., Li, J., and He, H. (2021). MS-MDA: multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition. *Front. Neurosci.* 15:778488. doi: 10.3389/fnins.2021.778488
- Chen, B., Lam, W., Tsang, I. W., and Wong, T. L. (2013). Discovering low-rank shared concept space for adapting text mining models. *IEEE Transac. Pattern Anal. Mach. Intell.* 35, 1284–1297. doi: 10.1109/TPAMI.2012.243
- Dan, Y., Tao, J., Fu, J., and Zhou, D. (2021). Possibilistic clustering-promoting semi-supervised learning for EEG-based emotion recognition. *Front. Neurosci. Brain Imag. Methods* 15:690044. doi: 10.3389/fnins.2021.690044
- Ding, Z., Nasser, M. N., and Fu, Y. (2018b). Semi-supervised deep domain adaptation via coupled neural networks. *IEEE Trans. Image Process.* 27, 5214–5224. doi: 10.1109/TIP.2018.2851067
- Ding, Z., Shao, M., and Fu, Y. (2018c). Incomplete multisource transfer learning. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 310–323. doi: 10.1109/TNNLS.2016.2618765
- Ding, Z., Sheng, L., Ming, S., and Fu, Y. (2018a). Graph adaptive knowledge transfer for unsupervised domain adaptation. 15th European Conference (ECCV 2018), Munich, Germany, September 8–14, 2018, Springer, Cham.
- Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194. doi: 10.1126/science.1076358
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2014). DeCAF: a deep convolutional activation feature for generic visual recognition. *PMLR* 32, 647–655. doi: 10.5555/3044805.3044879
- Du, X., Ma, C., Zhang, G., Li, J., Lai, Y. K., Zhao, G., et al. (2020). An efficient LSTM network for emotion recognition from multichannel EEG signals. *IEEE Trans. Affect. Comput.* 13, 1528–1540. doi: 10.1109/TAFFC.2020.3013711
- Duan, L., Tsang, I. W., and Xu, D. (2012). Domain transfer multiple kernel learning. *IEEE Transac. Pattern Anal. Mach. Intell.* 34, 465–479. doi: 10.1109/TPAMI.2011.114
- Fang, C., Xu, Y., and Rockmore, D. N. (2013). “Unbiased metric learning: on the utilization of multiple datasets and web images for softening bias” in *Proceedings of IEEE International Conference on Computer Vision*. 1657–1664.
- Feiping Nie, , Dong Xu, , Tsang, I. W. H., and Zhang, C. (2010). 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
- Gan, C., Yang, T., and Gong, B. (2016). “Learning attributes equals multisource domain generalization” in *Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR)*. 87–97.
- Ganin, Y., and Lempitsky, V. (2015). “Unsupervised domain adaptation by back-propagation” in *Proceedings of International Conference of Machine Learning (ICML)*, pp. 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, 2096–2030.
- Gao, J., Huang, R., and Li, H. (2015). Sub-domain adaptation learning methodology (SDAL). *Inf. Sci.* 298, 237–256. doi: 10.1016/j.ins.2014.11.041
- Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. (2017). Scatter component analysis: a unified framework for domain adaptation and domain generalization. *IEEE Transac. Pattern Anal. Mach. Intell.* 99:1. doi: 10.1109/TPAMI.2016.2599532
- Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. (2015). “Domain generalization for object recognition with multi-task autoencoders” in *Proceedings of IEEE International Conference of Computer Vision*. 2551–2559.
- Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. *Conference on Computer Vision and Pattern Recognition. IEEE*. 2066–2073.
- Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B.K. (2009). “A fast, consistent kernel two-sample test” in *Conference on Neural Information Processing Systems* 22, Vancouver, British Columbia, Canada. MIT Press. 673–681.
- Han, Z., Chang, H., Zhou, X., Wang, J., Wang, L., and Shao, Y. (2022). E2ENNet: an end-to-end neural network for emotional brain-computer interface. *Front. Comput. Neurosci.* 16:942979. doi: 10.3389/fncom.2022.942979
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778
- Jayaram, V., Alangir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Comput. Intell. Mag.* 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
- Judy, H., Kulis, B., Darrell, T., and Saenko, K. (2012). Discovering latent domains for multisource domain adaptation. *European Conference on Computer Vision*. Springer, Berlin, Heidelberg.
- Judy, H., Tzeng, E., Darrell, T., and Saenko, K. (2017). Simultaneous deep transfer across domains and tasks. *Domain Adapt. Comput. Vision Appl.* 2017, 173–187. doi: 10.1007/978-3-319-58347-1_9
- Kang, G., Jiang, L., Wei, Y., Yang, Y., and Hauptmann, A. Contrastive adaptation network for single- and multi-source domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1793–1804. doi: 10.1109/TPAMI.2020.3029948
- Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). “Undoing the damage of dataset bias” in *Proceedings of European Conference on Computer Vision*. 158–171.
- 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, 1–13. doi: 10.1155/2013/573734
- Koelstra, S., Muhl, 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 Muller-Putz, G. R. (2018). Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Transac. Cogn. Dev. Syst.* 11, 85–94. doi: 10.1109/TCDS.2018.2826840
- Li, H., Jin, Y.-M., Zheng, W.-L., and Lu, B. L. (2018). “Cross-subject emotion recognition using deep adaptation networks” in *Neural Information Processing*, eds. L. Cheng, A. C. S. Leung and S. Ozawa (Cham: Springer International Publishing), 403–413.
- Li, J., Qiu, S., Du, C., Wang, Y., and He, H. (2020). Domain adaptation for EEG emotion recognition based on latent representation similarity. *IEEE Transac. 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. (2019). Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Transac. Cybernet.* 50, 3281–3293. doi: 10.1109/TCYB.2019.2904052
- Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., and Hu, B. (2018). Exploring EEG features in cross-subject emotion recognition. *Front. Neurosci.* 12:162. doi: 10.3389/fnins.2018.00162
- Li, R., Wang, Y., and Lu, B. (2021). “A multi-domain adaptive graph convolutional network for EEG-based emotion recognition” in *Proceedings of the 29th ACM International Conference on Multimedia (MM’21)*, October 20–24. Virtual Event, China. ACM, New York, NY, USA, p. 9.
- Li, W., Xu, Z., Xu, D., Dai, D., and van Gool, L. (2018). Domain generalization and adaptation using low-rank exemplar SVMs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 1114–1127. doi: 10.1109/TPAMI.2017.2704624
- 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 *The 27th International Joint Conference on Artificial Intelligence (IJCAI)*.
- Li, Y., Zheng, W., Cui, Z., Zong, Y., and Ge, S. (2018a). EEG emotion recognition based on graph regularized sparse linear regression. *Neural. Process. Lett.* 47, 1–19. doi: 10.1007/s11063-017-9609-3
- Li, Y., Zheng, W., Wang, L., Zong, Y., and Cui, Z. (2019). From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* 13, 568–578. doi: 10.1109/TAFFC.2019.2922912
- Li, Y., Zheng, W., Zong, Y., Cui, Z., Zhang, T., and Zhou, X. (2018c). A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* 12, 494–504. doi: 10.1109/TAFFC.2018.2885474
- Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). “Learning transferable features with deep adaptation networks” in *Proceedings of International Conference on Machine Learning*. 97–105.
- Long, M., Cao, Z., Wang, J., and Jordan, M. I. (2018). “Conditional adversarial domain adaptation” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 1647–1657.

- Long, M., Wang, J., Ding, G., Pan, S. J., and Yu, P. S. (2014b). Adaptation regularization: a general framework for transfer learning. *IEEE Transac. Knowledge Data Eng.* 26, 1076–1089. doi: 10.1109/TKDE.2013.111
- Long, M., Wang, J., Ding, G., Sun, J., and Philip, S.Y. (2014a). “Transfer joint matching for unsupervised domain adaptation” in *Conference on Computer Vision and Pattern Recognition*. IEEE. 1410–1417.
- Long, M., Wang, J., and Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. *Proc. Int. Conf. Mach. Learn.* 70, 2208–2217. doi: 10.5555/3305890.3305909
- Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2016). “Unsupervised domain adaptation with residual transfer networks” in *Proceedings of the Neural Information Processing Systems*. 136–144.
- Luo, Y., Zhang, S. Y., Zheng, W. L., and Lu, B.-L. (2018). “WGAN domain adaptation for EEG-based emotion recognition” in *International Conference on Neural Information Processing*.
- Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). “Domain adaptation with multiple sources” in *Conference on Neural Information Processing Systems*. Vancouver, British Columbia, Canada, MIT Press. 1041–1048.
- Motiian, S., Piccirilli, M., Adjero, D. A., and Doretto, G. (2017). “Unified deep supervised domain adaptation and generalization” in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy. pp. 5716–5726.
- Muandet, K., Baldazzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. *Proc. Int. Conf. Mach. Learn.* 28, 10–18.
- Mughal, N. E., Khan, M. J., Khalil, K., Javed, K., Sajid, H., Naseer, N., et al. (2022). EEG-fNIRS-based hybrid image construction and classification using CNN-LSTM. *Front. Neurobot.* 16:873239. doi: 10.3389/fnbot.2022.873239
- 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. Interf.* 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. (2010). “Efficient and robust feature selection via joint-norms minimization” in *International Conference on Neural Information Processing Systems*. Curran Associates Inc. 1813–1821
- Niu, L., Li, W., and Xu, D. (2015). “Visual recognition by learning from web data: a weakly supervised domain generalization approach” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2774–2783.
- Niu, L., Li, W., Xu, D., and Cai, J. (2018). An exemplar-based multi-view domain generalization framework for visual recognition. *IEEE Transac. Neural Netw. Learn. Syst.* 29, 259–272. doi: 10.1109/TNNLS.2016.2615469
- 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 *Lecture Notes on Data Engineering and Communications Technologies, Data Science and Big Data Analytics* (Springer), 117–124.
- Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation: a survey of recent advances. *IEEE Signal Process. Mag.* 32, 53–69. doi: 10.1109/MSP.2014.2347059
- Pei, Z., Cao, Z., Long, M., and Wang, J. (2018). Multi-adversarial domain adaptation. *Proc. AAAI* 32, 3934–3941. doi: 10.1609/aaai.v32i1.11767
- Rosenstein, M.T., Marx, Z., and Kaelbling, L. P. “To transfer or not to transfer” in *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” in *The 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka. 6627–6630.
- Simonyan, K., and Zisserman, A. (2014). “Very deep convolutional networks for large-scale image recognition” in *Proceedings of International Conference Learning Representations (ICLR)*. 1–14.
- Song, T., Zheng, W., Song, P., and Cui, Z. (2018). “EEG emotion recognition using dynamical graph convolutional neural networks” in *IEEE Transactions on Affective Computing*, pp. 1.
- Sun, B., and Saenko, K. (2016). Deep CORAL: correlation alignment for deep domain adaptation. *Proc. ECCV Workshops*.
- 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., 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
- Tao, J., Di, Z., Fangyu, L., and Bin, Z. (2019). Latent multi-feature co-regression for visual recognition by discriminatively leveraging multi-source models. *Pattern Recogn.* 87, 296–316. doi: 10.1016/j.patcog.2018.10.023
- Tao, J., 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 Cate FOries from few examples with multi model knowledge transfer [J]. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 928–941. doi: 10.1109/TPAMI.2013.197
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). “Adversarial discriminative domain adaptation” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA. 2962–2971.
- Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. arXiv [Preprint]. doi: 10.48550/arXiv.1412.3474
- Wang, C., and Mahadevan, S. (2011). Heterogeneous domain adaptation using manifold alignment. *International Joint Conference on Artificial Intelligence*, Barcelona, Catalonia, Spain. 1541–1546.
- Wang, Y., Qiu, S., Li, D., Du, C., Lu, B. L., and He, H. (2022). Multi-modal domain adaptation variational autoencoder for EEG-based emotion recognition. *IEEE/CAA J. Automat. Sin.* 9, 1612–1626. doi: 10.1109/JAS.2022.105515
- Xiao, M., and Guo, Y. (2015). Feature space independent semi-supervised domain adaptation via kernel matching. *IEEE Transac. Pattern Anal. Mach. Intellig.* 37, 54–66. doi: 10.1109/TPAMI.2014.2343216
- Xu, Y., He, X., Xu, G., Qi, G., Yu, K., Yin, L., et al. (2022). A medical image segmentation method based on multi-dimensional statistical features. *Front. Neurosci.* 16:1009581. doi: 10.3389/fnins.2022.1009581
- Xu, Z., Li, W., Niu, L., and Xu, D. (2014). “Exploiting low-rank structure from latent domains for domain generalization” in *Proceedings of European Conference on Computer Vision*. 628–643
- 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 Transac. Pattern Anal. Mach. Intellig.* 29, 40–51. doi: 10.1109/TPAMI.2007.250598
- Yang, Y., Ma, Z., Hauptmann, A. G., and Sebe, N. (2013). Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Trans. Multimed.* 15, 661–669. doi: 10.1109/TMM.2012.2237023
- Yang, J., Yan, R., and Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive SVMs. *ACM Int. Conf. Multimedia ACM* 2007, 188–197. doi: 10.1145/1291233.1291276
- Zhang, Y., Chen, J., Tan, J. H., Chen, Y., Chen, Y., Li, D., et al. (2020c). An investigation of deep learning models for EEG-based emotion recognition. *Front. Neurosci.* 14:622759. doi: 10.3389/fnins.2020.622759
- 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: 10.3233/JIFS-182561
- 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.knsys.2019.105330
- 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. Humaniz. Comput.* 2019, 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. Imag. Health Inform.* 7, 1654–1659. doi: 10.1166/jmihi.2017.2181
- 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. Imag. 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 Transac. 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. Auton. 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 Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 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 *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.* 13, 1290–1301. doi: 10.1109/TAFFC.2020.2994159

Zhong, Q., Zhu, Y., Cai, D., Xiao, L., and Zhang, H. (2020). Electroencephalogram Access for emotion recognition based on a deep hybrid network. *Front. Hum. Neurosci.* 14:589001. doi: 10.3389/fnhum.2020.589001

Zhou, R., Zhang, Z., Fu, H., Zhang, L., Li, L., Huang, G., et al. (2022). PR-PL: a novel transfer learning framework with prototypical representation based pairwise learning for EEG-based emotion recognition. arXiv [Preprint]. doi: 10.48550/arXiv.2202.06509

Zhu, Y., Zhuang, F., Wang, J., Ke, G., Chen, J., Bian, J., et al. (2020). Deep subdomain adaptation network for image classification. *IEEE Transac. Neural Netw. Learn. Syst.* 99, 1–10. doi: 10.1109/TNNLS.2020.2988928