frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 23 July 2025
DOI 10.3389/fnins.2025.1592070

OPEN ACCESS

EDITED BY
Angarai Ganesan Ramakrishnan,
Indian Institute of Technology Hyderabad, India

REVIEWED BY
Jerrin Thomas Panachakel,
College of Engineering, Trivandrum, India
Anoop C. S.,
Government Engineering College Kozhikode, India

*CORRESPONDENCE
Di Zhou
✉ sion2005@sasu.edu.cn

RECEIVED 12 March 2025
ACCEPTED 16 June 2025
PUBLISHED 23 July 2025

CITATION
Dan Y, Li Q, Wang X and Zhou D (2025) Domain adaptive deep possibilistic clustering for EEG-based emotion recognition. *Front. Neurosci.* 19:1592070. doi: 10.3389/fnins.2025.1592070

Domain adaptive deep possibilistic clustering for EEG-based emotion recognition

Yufang Dan[1,2,3,4,5], Qun Li[1], Xianhua Wang[1] and Di Zhou[6]*

[1]Institute of Artificial Intelligence Application, Ningbo Polytechnic, Ningbo, China, [2]Ningbo Key Laboratory of Aging Health Equipment and Service Technology, Ningbo, China, [3]Dazhou City Key Laboratory of Multidimensional Data Perception and Intelligent Information Processing, Dazhou, Sichuan, China, [4]Sichuan Provincial Engineering Research Center of Meteorological Photoelectric Sensing Technology and Applications, Dazhou, Sichuan, China, [5]Special Polymer Materials for Automobile Key Laboratory of Sichuan Province, Sichuan University of Arts and Science, Dazhou, China, [6]Industrial Technological Institute of Intelligent Manufacturing, Sichuan University of Arts and Science, Sichuan, China

Emotion recognition based on electroencephalogram (EEG) faces substantial challenges. The variability of neural signals among different subjects and the scarcity of labeled data pose obstacles to the generalization ability of traditional domain adaptation (DA) methods. Existing approaches, especially those relying on the maximum mean discrepancy (MMD) technique, are often highly sensitive to domain mean shifts induced by noise. To overcome these limitations, a novel framework named **D**omain **A**daptive **D**eep **P**ossibilistic **c**lustering (**DADPc**) is proposed. This framework integrates deep domain-invariant feature learning with possibilistic clustering, reformulating Maximum Mean Discrepancy (MMD) as a one-centroid clustering task under a fuzzy entropy-regularized framework. Moreover, the DADPc incorporates adaptive weighted loss and memory bank strategies to enhance the reliability of pseudo-labels and cross-domain alignment. The proposed framework effectively mitigates noise-induced domain shifts while maintaining feature discriminability, offering a robust solution for EEG-based emotion recognition in practical applications. Extensive experiments conducted on three benchmark datasets (SEED, SEED-IV, and DEAP) demonstrate the superior performance of DADPc in emotion recognition tasks. The results show significant improvements in recognition accuracy and generalization capability across different experimental protocols, including cross-subject and cross-session scenarios. This research contributes to the field by providing a comprehensive approach that combines deep learning with possibilistic clustering, advancing the state-of-the-art in cross-domain EEG analysis.

KEYWORDS

electroencephalography, emotion recognition, deep domain adaptation, clustering assumption, memory bank

1 Introduction

In the field of affective computing (Muhl C. and G., 2014), automatic emotion recognition (AER) (Stern, 2002) has gained significant attention (Kim et al., 2013; Zhang et al., 2013), especially for EEG-based emotion recognition (Wenming, 2017; Li et al., 2018a; Pandey and Seeja, 2019; Jenke et al., 2014; Musha T. and A., 1997). From a machine learning perspective, EEG-based AER tasks are typically formulated as classification or regression problems (Kim et al., 2013; Zhang et al., 2013). However, due to inter-subject variability in emotional expression patterns (Pandey and Seeja, 2019), classifiers trained on specific subjects often have poor generalization ability. Although optimizing feature

representations and learning models has improved recognition accuracy (Li et al., 2018, 2019; Du et al., 2022; Song et al., 2018; Zhong et al., 2020; Zheng and Lu, 2015; Wei et al., 2015; Zhou et al., 2023), applying these classifiers to new subjects still yields unsatisfactory results (Zheng and Lu, 2013; Ghifary et al., 2017; Lan et al., 2019; Vinay et al., 2015; Wang et al., 2023). Domain Adaptation (DA) has emerged as a solution, aiming to transfer knowledge from related source domains to the target domain with scarce labeled data (VM Patel, 2015; Dan et al., 2022; Tao et al., 2017; Zhang et al., 2019; Tao et al., 2022).

The key to effective knowledge transfer in DA is to ensure data distribution similarity between the source domain and the target domain. Existing DA approaches mainly focus on distribution matching (such as instance re-weighting and feature mapping) and classifier model adaptation (Pan and Yang, 2018; VM Patel, 2015; Pan et al., 2011; Gretton et al., 2009; Chu et al., 2013; Long et al., 2013; Mahsa et al., 2013; Ganin et al., 2016; Kang et al., 2022; Liang et al., 2019; Tao et al., 2019, 2012, 2015, 2016). Early methods for addressing domain distribution shift, such as instance weighting techniques, including the popular Maximum Mean Discrepancy (MMD) (Gretton et al., 2009), have limitations. MMD often decouples optimization from classifier training. Feature mapping approaches (Pan et al., 2011; Long et al., 2013; Mahsa et al., 2013; Kang et al., 2022), like Transfer Component Analysis (TCA) (Pan et al., 2011) and Joint Domain Adaptation (JDA) (Long et al., 2013), have been developed to address these issues, but they still have drawbacks. Mahsa et al. (2013) introduced the Domain Invariant Projection (DIP) algorithm, which employs a polynomial kernel on the MMD metric to establish a concise shared feature space and decrease intra-class dispersion through a clustering-based method.

Conventional MMD-based DA approaches overlook the statistical framework of the target domain, which can impede accurate label prediction. Some methods, including the Contrastive Adaptation Network and Domain Invariant Projection Ensemble (Kang et al., 2022) attempt to address this issue, yet they remain MMD-based. Moreover, current MMD-based methods do not fully account for intra-domain noise, which can lead to mean-shift issues and compromise generalization.

Possibilistic clustering frameworks (Dan et al., 2024; Krishnapuram and Keller, 1993) offer a solution to these problems as they can mitigate noise interference during data clustering. The conventional MMD metric has been adapted into a single-cluster center objective in a noisy context, and a resilient domain adaptation classifier (EDPC) (Dan et al., 2024) based on the possibilistic distribution distance metric has been proposed. However, EDPC, as a shallow DA method, has limited feature extraction capabilities. Deep neural networks, with their powerful feature-extraction ability, have led to the development of deep DA models (Long et al., 2015; Mingsheng Long and Wang Jianmin, 2016; Chen et al., 2019; Lee et al., 2019; Ding et al., 2018; Tang and Jia, 2019). Contemporary affective models often use deep transfer learning methods like domain-adversarial neural networks (DANN) (Ganin et al., 2016). Although these models can reduce domain distribution differences in large datasets, they have not fully resolved the domain-shift problem with small datasets(i.e., EEG datasets).

In this study, we propose a novel Domain Adaptive Deep Possibilistic clustering (DADPc) approach for EEG-based emotion recognition. It integrates an adaptive loss function with a fuzzy entropy regularization mechanism to enhance the model's cross-domain adaptability and clustering performance. The main contributions are as follows:

- Integrating clustering with neural network training, creating a DADPc criterion for simultaneous feature reconstruction and clustering based on deep features.
- Using a robust loss function with adaptive weights and fuzzy entropy increases insensitivity to outliers and introduces fuzzy entropy regularization for the affinity matrix. The affinity and possibilistic centroid matrices are updated efficiently without using stochastic gradient descent.
- Demonstrating the effectiveness of the proposed method through extensive experiments on multiple EEG datasets (SEED, SEED-IV, and DEAP).

2 Related research

Over the past decade, research on using EEG signals for emotion recognition has burgeoned (Shi et al., 2013; Zheng and Lu, 2015; Li et al., 2016; Alarco and Fonseca, 2019; Luo et al., 2024; Zhong et al., 2020; Chen et al., 2021; Zheng and Lu, 2013). Early on, Shi et al. (2013) employed EEG features and SVMs for emotion classification. With advancements in deep learning, deep neural networks gained popularity in EEG-based emotion recognition (Luo et al., 2024; Zhou et al., 2023; Zhong et al., 2020). These models excel in subject-dependent tasks but struggle with subject-independent tasks due to inter-subject EEG variability (Luo et al., 2024; Zhou et al., 2023; Zheng and Lu, 2013).

To address this issue, transfer learning strategies have been implemented. Zheng and Lu (2013) utilized non-deep transfer learning methods such as TCA (Pan et al., 2011) and TPT for cross-subject emotion recognition. Jin et al. (2017) introduced DANN-based deep transfer learning for EEG emotion recognition, outperforming non-deep methods. Following this development, refined DANN-based architectures have emerged (Li et al., 2019b; Peng et al., 2022). For example, Li et al. (2019b) minimized distribution divergence, while Peng et al. (2022) proposed JAGP, both of which enhance cross-subject performance. However, deep transfer learning has limitations (Luo et al., 2024) regarding label noise or small datasets.

Given these challenges, new transfer learning techniques and shallow methods are needed, especially for using labeled source data alongside unlabeled data augmented by pseudo-labels (Peng et al., 2023, 2022; Magdiel and Gibran, 2023; Zhou et al., 2023; Tao and Dan, 2021; Tao et al., 2023; Dan et al., 2022). Zhang and Etemad (2022) introduced the PARSE model to address domain distribution mismatch in semi-supervised EEG emotion recognition, enhancing cross-subject performance. Dan et al. (2024) proposed a semi-supervised model with possibilistic clustering, which requires less labeled data.

To improve the accuracy of shallow methods, we explore a possibilistic clustering method with deep learning features to

develop a classifier. Our aim is to address the challenges posed by noise and the limitations of small-scale datasets while enhancing the accuracy of emotion recognition.

3 Preliminary

In this study, matrices are denoted by uppercase bold letters, while vectors are represented by lowercase bold letters. For a given matrix $\mathbf{Q}$, $\mathbf{q}^i$ refers to the i^{th} row vector, $\mathbf{q}_i$ represents the i^{th} column vector, and q_{ij} denotes the element at position (i, j). Additionally, $\mathbf{Q}^T$ signifies the transpose of $\mathbf{Q}$. The vector $\mathbf{1} = [1, 1, \cdots, 1]^T$ consists entirely of ones, while $\mathbf{0}$ denotes a zero matrix. The notation $\mathbf{Q} > 0$ indicates that all elements of the matrix are positive. The ℓ_1-norm and ℓ_2-norm of a vector $\mathbf{q}$ are expressed as $\|\mathbf{q}\|_1$ and $\|\mathbf{q}\|_2$, respectively. For a scalar-output function $f(\mathbf{x})$, the gradient is given by $\nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right]^T$. For a vector-output function $\boldsymbol{g}(x)$, the gradient with respect to x is denoted as $\nabla_x \boldsymbol{g}(x) = \left[\frac{\partial g_1}{\partial x}, \frac{\partial g_2}{\partial x}, \cdots, \frac{\partial g_n}{\partial x}\right]$.

In DA learning, the source domain is defined as $\mathcal{D}_{\mathcal{S}} = \left\{x_i^s, y_i^s\right\}_{i=1}^{n_s}$, where the sample set is defined as $X^s = \left[x_1^s, \ldots, x_{n_s}^s\right] \in \mathbb{R}^{d \times n_s}$, and the corresponding class labels are defined as $Y^s = \left\{y_1, \ldots, y_{n_s}\right\}^T \in \{0, 1\}^{n_s \times K}$. d is the sample(i.e., x_i) dimension of the source domain. n_s is the sample number of the source domain. K is the class number of the source domain. Here, $y_i \in \{0, 1\}^{K \times 1}$ is a one-hot encoded vector; if x_i belongs to the j-th class, then $y_{ij} = 1$, the rest of the elements of y_i are 0. The unlabeled target domain is defined as $\mathcal{D}_{\mathcal{T}} = \left\{x_j^t\right\}_{j=1}^{n_t}$, where the sample set and unknown sample labels during training are $X^t = \left[x_1^t, \ldots, x_{n_t}^t\right] \in \mathbb{R}^{d \times n_t}$, $Y^t = \left\{y_1, \ldots, y_{n_t}\right\}^T \in \mathbb{R}^{n_t \times K}$, respectively. n_t is the sample number of target domain. We further define $X = \left[X^s, X^t\right] \in \mathbb{R}^{d \times N}$ and $Y = \left[Y^s, Y^t\right] \in \mathbb{R}^{N \times K}$, $N = n_s + n_t$.

3.1 Adaptive loss function

For any given vector $\mathbf{q}$, the ℓ_1-norm and the squared ℓ_2-norm are defined as $\|\mathbf{q}\|_1 = \sum_i |q_i|$ and $\|\mathbf{q}\|_2^2 = \sum_i q_i^2$, respectively. To leverage the benefits of different norms, a robust loss function, known as the adaptive loss function, is defined as Nie et al. (2013):

$$\|\mathbf{Q}\|_\sigma = \sum_i \frac{(1+\sigma)\left\|\mathbf{q}^i\right\|_2^2}{\left\|\mathbf{q}^i\right\|_2 + \sigma} \tag{1}$$

, where σ serves as a trade-off parameter that governs robustness to different types of outliers. The properties of $\|\mathbf{Q}\|_\sigma$ are summarized in Nie et al. (2013).

3.2 Possibilistic clustering

In a specific reproducing kernel Hilbert space (RKHS), denoted as $\mathcal{H}$, the data from the original space can undergo a non-linear transformation ϕ that maps it into a feature representation within the RKHS (Mingsheng Long and Wang Jianmin, 2016). This transformation is of the form $\phi : \mathbb{R}^d \to \mathcal{H}$. The associated kernel function, designated as $K(..) : X \times X \to \mathbb{R}$, is defined by $K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}$ for $x_1, x_2 \in X$. This kernel technique is widely utilized in contemporary non-linear learning approaches (Pan et al., 2011; Long et al., 2015). Research has demonstrated (Gretton et al., 2009; Bruzzone and Marconcini, 2010) that mapping sample data to a high- or infinite-dimensional space can facilitate the capture of higher-dimensional data features (Carlucci et al., 2017). In an RKHS, the maximum mean discrepancy (MMD) criterion effectively gauges the distance between two distributions. Accordingly, let $\mathcal{F}$ represent a set of functions of a particular kind, $f : \mathcal{X} \to \mathbb{R}$. The MMD between two domain distributions, $\mathbb{P}$ and $\mathbb{Q}$, is defined as follows:

$$MMD_{\mathcal{F}}[\mathbb{P}, \mathbb{Q}] := \sup_{f \in \mathcal{F}} \left(\mathop{\mathbb{E}}_{\mathbb{P}}[f(x)] - \mathop{\mathbb{E}}_{\mathbb{Q}}[f(x)] \right). \tag{2}$$

The maximum mean discrepancy (MMD) metric aims to reduce the anticipated disparity between two domain distributions using a specific function f, thereby maximizing their similarity. As the size of the domain sample becomes sufficiently large (or approaches infinity), the anticipated disparity converges to (or matches) the empirical mean difference. Consequently, Equation 3 can be expressed in terms of the empirical MMD form.

$$MMD\left(X^s, X^t\right) := \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi\left(x_i^s\right) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi\left(x_j^t\right) \right\|_M^2. \tag{3}$$

To establish the universal link between the conventional Maximum Mean Discrepancy (MMD) criterion and the clustering model based on means, we present the theorem below:

Theorem 1. The Maximum Mean Discrepancy (MMD) metric can be framed approximately as a unique type of clustering task with a single cluster centroid, denoted by μ, where the assignment of instances to the cluster is represented by ζ_k.

$$MMD\left(X^s, X^t\right) \leq \sum_{k=1}^{N} \zeta_k \|\phi(x_k) - \mu\|_H^2 \tag{4}$$

Proof.

$$\begin{aligned}
&MMD\left(X^s, X^t\right) \\
&= \left\| \tfrac{1}{n_s} \textstyle\sum_{i=1}^{n_s} x_i^s - \tfrac{1}{n_t} \sum_{j=1}^{n_t} x_j^t \right\|_H^2 \\
&= \left\| \tfrac{1}{n_s} \textstyle\sum_{i=1}^{n_s} x_i^s - \mu + \mu - \tfrac{1}{n_t} \sum_{j=1}^{n_t} x_j^t \right\|_H^2 \\
&\leq \left\| \tfrac{1}{n_s} \textstyle\sum_{i=1}^{n_s} x_i^s - \mu \right\|_H^2 + \left\| \tfrac{1}{n_t} \textstyle\sum_{j=1}^{n_t} x_j^t - \mu \right\|_H^2 \\
&= \tfrac{1}{n_s^2} \left\| \textstyle\sum_{i=1}^{n_s} x_i^s - n_s \mu \right\|_H^2 + \tfrac{1}{n_t^2} \left\| \textstyle\sum_{j=1}^{n_t} x_j^t - n_t \mu \right\|_H^2 \\
&= \tfrac{1}{n_s^2} \left\| \textstyle\sum_{i=1}^{n_s} \left(x_i^s - \mu\right) \right\|_H^2 + \tfrac{1}{n_t^2} \left\| \textstyle\sum_{j=1}^{n_t} \left(x_j^t - \mu\right) \right\|_H^2 \\
&\leq \tfrac{1}{n^2} \textstyle\sum_{i=1}^{n_s} \left\| x_i^s - \mu \right\|_H^2 + \tfrac{1}{n_t^2} \sum_{j=1}^{n_t} \left\| x_j^t - \mu \right\|_H^2 \\
&= \textstyle\sum_{k=1}^{N} \zeta_k \|x_k - \mu\|_H^2,
\end{aligned} \tag{5}$$

where $\mu = \delta\mu + (1 - \delta)\mu_t$ is the cluster center with $0 \leq \delta \leq 1$. μ_s and μ_t are the means of the source domain and the target domain, respectively. When $n_s = n_t$, let $\delta = 0.5$. When $n_s \neq n_t$, the number

of data in the source domain and target domain can be set the same during sampling. The sample membership ζ_k is defined as follows:

$$\zeta_k = \begin{cases} \frac{1}{n_s^2}, & x_k \in X^s \\ \frac{1}{n_t^2}, & x_k \in X^t \end{cases} \tag{6}$$

Theorem 1 highlights the intrinsic connection between the MMD criterion for domain distribution and the clustering model. This connection can facilitate more efficient alignment of distributions across distinct domains through the clustering of domain data. However, it is important to note that traditional clustering models are susceptible to noise, as pointed out by by Dan et al. (2024). Consequently, DA methods relying on MMD often face the challenge of domain mean shift due to noisy data. To tackle this problem, this paper explores more robust clustering approaches and introduces a novel, effective criterion for measuring domain distribution distance in the following section.

4 Methodology

Traditional methods such as Kernel K-Means (KKM) and Possibilistic clustering are impractical for large-scale datasets, while efficient algorithms like K-means and Possibilistic clustering are overly simplistic for non-linear data. To address this issue, we propose the DADPc model, which performs effectively on both large datasets and non-linearly distributed data. This section begins by introducing the adaptive loss function and entropy regularization to enhance possibilistic clustering. The architecture of our proposed method is depicted in Figure 1. Our method efficiently utilizes concurrent deep features from the source and target domains for domain adaptation, based on the general framework outlined in Section 4.1. In particular, we enhance the sample reconstruction process using an encoder-decoder deep neural network, as described in Section 4.2, and present our technique for generating the source domain with an adaptive loss function in Section 4.3. Additionally, we introduce a deep clustering approach that incorporates a memory bank and possibilistic theory in Section 4.4. This approach enables us to efficiently extract the essential features from both domains while guaranteeing the cross-domain transferability of the acquired features. Subsequently, the specifics of the DADPc model are discussed in the following subsections.

4.1 General formulation

For the problem of DA in complex structures and noisy environments, we aim to improve the robustness of distribution distance metrics for DA and enhance generalization in the target domain. Based on the DA generalization error theory (Ben-David et al., 2010), we seek to achieve the following two core objectives: First, we construct a robust distribution distance metric that can resist the impact of noise, addressing the issue of domain mean-shift. The differences in domain distribution can be selectively corrected. Second, we effectively perform semantic reasoning in the target domain by maintaining the geometric structure consistency of the data in the domain, connecting the discriminative information of the source domain, and minimizing the discriminative error in the target domain. A highly generalizable target domain classifier is constructed. Therefore, our general framework can be described as follows:

$$\min_{\Gamma} \mathcal{J} = \min_{\mathbf{W}^{(m)},\mathbf{b}^{(m)},\mathbf{V},\mathbf{C}} \mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3 \tag{7}$$

, where $\Gamma = \left\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}, \mathbf{V}, \mathbf{C}\right\}$, $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$ are designed for different purposes. $\mathcal{J}_1$ ensures the minimum reconstruction error based on data from both domains. $\mathcal{J}_2$ is the classification error for the source domain and a regularization used to avoid the overfitting of auto-encoders, and also able to prevent the auto-encoder from generating a trivial map. $\mathcal{J}_3$ is the cost function of deep possibilistic clustering with a memory bank. Inspired by the adaptive loss function outlined in Equation 1, which serves as an interpolation between the $\ell_{2,1}$-norm and the Frobenius norm squared, we introduce a new objective function for feature reconstruction, the parameters (i.e., W and b) of deep neural networks, and deep possibilistic clustering with an adaptive loss in $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$.

4.2 Feature reconstruction

The DADPc employs a neural network architecture comprising $(M+1)$ layers to transform raw data into a nonlinear feature space, where M is an even integer. The initial $\frac{M}{2}$ hidden layers function as an encoder, responsible for reducing the dimensionality of the input data. Conversely, the remaining $\frac{M}{2}$ hidden layers act as a decoder, tasked with the reconstruction of the data. Given that $\mathbf{H}^{(0)} = \mathbf{X} \in {}^{d\times N}$, where $\mathbf{X}$ belongs to the union of source and target datasets $X_s \cup X_t$, and N equals the sum $n_s + n_t$ representing the data counts in each domain, $\mathbf{h}_i^{(0)}$ signifies the i-th column vector of $\mathbf{H}^{(0)}$, equivalent to $\mathbf{x}_i$ and $\mathbf{x}_i \in \mathbf{X}$. The output generated by the m-th layer is denoted as follows:

$$\mathbf{h}_i^{(m)} = f\left(\mathbf{W}^{(m)}\mathbf{h}_i^{(m-1)} + \mathbf{b}^{(m)}\right) \in \mathbb{R}^{d_m} \tag{8}$$

where m ranges from 1 to M, d_m represents the number of neurons in the m-th layer. The activation function for each layer is denoted by $f(\cdot)$. The weight matrix and bias for the respective layer, as indicated in Equation 8, are given by $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$. When presented with a data point $\mathbf{x}_i$, the auto-encoder network initially maps the raw data onto a non-linear, low-dimensional space, which can be expressed as follows:

$$\mathbf{H}^{(\frac{M}{2})} = \left[\mathbf{h}_1^{(\frac{M}{2})}, \mathbf{h}_2^{(\frac{M}{2})}, \cdots, \mathbf{h}_n^{(\frac{M}{2})}\right] \in \mathbb{R}^{d_{\frac{M}{2}}\times N} \tag{9}$$

then reconstructs the feature as:

$$\mathbf{H}^{(M)} = \left[\mathbf{h}_1^{(M)}, \mathbf{h}_2^{(M)}, \cdots, \mathbf{h}_n^{(M)}\right] \in \mathbb{R}^{d\times N} \tag{10}$$

We employ deep neural networks to reconstruct the features from the source domain and the target domain, which encompass deep learning, meticulous analysis, and the effective transformation

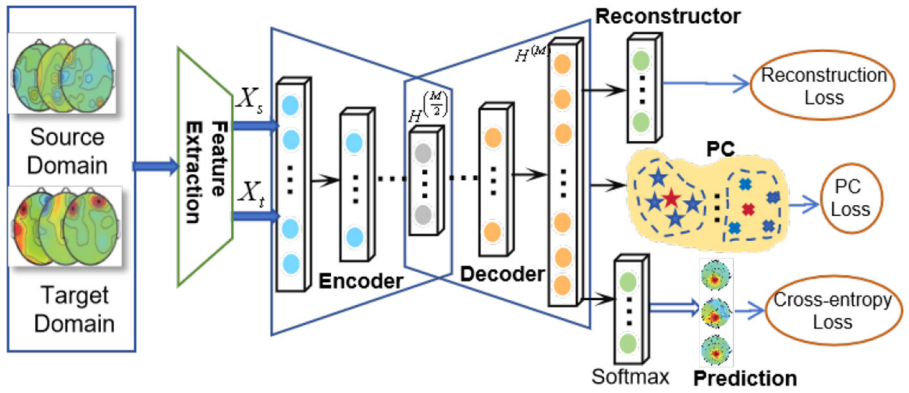

FIGURE 1
The architecture of DADPc: X_s and X_t [i.e., $H^{(0)}$] are inputs of the auto-encoder, $H^{(M)}$ is the output of the auto-encoder, i.e., the reconstruction of raw data, and $H^{(\frac{M}{2})}$ represents the deep features.

of both domain features. It aims to generate new features that are highly similar to the original or possess specific attributes.

$$\mathcal{J}_1 = \min_{\mathbf{W}^{(m)},\mathbf{b}^{(m)}} \left\| \mathbf{H}^{(M)} - \mathbf{X} \right\|_\sigma^2, \tag{11}$$

, where $\mathbf{X} = \cup_{k=1\ldots K}\mathbf{X}_{c_k}$, $\mathbf{X}_{c_k}$ includes all x_i which belong to class k, and $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$.

By leveraging data reconstruction techniques, the model captures the intrinsic characteristics and structural information of the data, thereby extracting more expressive features while filtering out noise and optimizing data quality. This series of operations enhances the robustness and generalization capabilities of subsequent classification and clustering tasks.

4.3 Source classifier

To learn a deep source classification model $h_s : \mathcal{X}_s \rightarrow \mathcal{Y}_s$, we aim to minimize the cross-entropy loss $\mathcal{J}_2$

$$\begin{aligned}\mathcal{J}_2 = &-\sum_{i=1}^{n_s}\sum_{k=1}^{K} q_k \log p_k(x_i^s) + \lambda_1 \sum_{m=1}^{M} \left\| \mathbf{W}^{(m)} \right\|_\sigma^2 \\ &+\lambda_2 \sum_{m=1}^{M} \left\| \mathbf{b}^{(m)} \right\|_{\sigma 1}^2,\end{aligned} \tag{12}$$

where $\mathbf{x}_i \in \mathbf{X}_{c_k}^s$, q denotes a one-hot encoding of y_i^s where q_k is "1" for the correct class and "0" for the rest. The last two terms are regularization used to avoid the overfitting of auto-encoders with regularization parameters λ_1 and λ_2. The two terms also prevent the auto-encoder from generating a trivial map. $p_k(x_i) = \frac{\exp\left(h_{ik}^{(M)}\right)}{\sum_{j=1}^{K}\exp\left(h_{ij}^{(M)}\right)}$ represents the predicted probability that sample x_i belongs to class k. Before applying the cross-entropy loss to the predictions of the source domain classifier, we first use the sharpening technique to address the ambiguity in the predictions of the source domain data:

$$\tilde{p}_k(x_i^s) = p_k(x_i^s)^{-\tau} / \sum_{e=1}^{K} p_e(x_i^s)^{-\tau}. \tag{13}$$

where τ represents the temperature parameter utilized for scaling prediction probabilities. When τ approaches 0, the probability distribution converges to a single point mass (Lee, 2013). Thus, Equation 12 can be reformulated as follows:

$$\begin{aligned}\mathcal{J}_2 = &-\sum_{i=1}^{n_s}\sum_{k=1}^{K} q_k \log \tilde{p}_k(x_i^s) + \lambda_1 \sum_{m=1}^{M} \left\| \mathbf{W}^{(m)} \right\|_\sigma^2 \\ &+\lambda_2 \sum_{m=1}^{M} \left\| \mathbf{b}^{(m)} \right\|_{\sigma_1}^2,\end{aligned} \tag{14}$$

4.4 Deep possibilistic clustering with memory Bank

Recent studies have shown the efficacy of probabilistic clustering methods in mitigating the adverse effects of noise on clustering outcomes (Dan et al., 2021). Consequently, this section generalizes the initial one-cluster center method (Theorem 1) to the context of deep probabilistic one-cluster centering. Subsequently, we introduce a distance metric for deep possibilistic clustering distributions, termed DPC. By incorporating the concept of deep possibilistic clustering entropy, we extend the rigid clustering method of MMD to a more flexible clustering framework. In this framework, the contribution of each sample is weighted based on its proximity to the overall domain mean: data farther from the mean contribute less and are more likely to be viewed as noise. Thus, DPC modulates the influence of noise-induced mean shift during domain alignment. The formula for the deep possibilistic clustering distribution distance metric at the M^{th} layer is defined as follows:

$$\mathcal{J}_3 = \sum_{i=1}^{N}\sum_{k=1}^{K} v_{i,k}^2 \left\| \mathbf{h}_i^{(M)} - \mathbf{c}_k \right\|_\sigma \tag{15}$$

The k-th cluster centroid is denoted by $\mathbf{c}_k \in \mathbb{R}^{d\times 1}$. The initial cluster centroids are derived from the source domain, as it provides labels for each sample. The cluster centroid for the emotion category k can be calculated by averaging all sample

features that belong to this category, expressed as follows:

$$c_k = \frac{1}{\left|X^s_{c_k}\right|} \sum_{x_i^s \in X^s_{c_k}} h_i^{(M)} \tag{16}$$

$v_{i,k}$ represents the possibility membership of the feature $h_i^{(M)}$, extracted from the sample x_i in the M layer of the Encoder-Decoder, belonging to the k-th class.

To enhance the resilience and efficacy of the possibilistic clustering method for measuring distribution distance in noisy datasets, a regularization term involving fuzzy entropy is introduced in Equation 17. This term is associated with the parameter $v_{i,k}$:

$$\mathcal{J}_3 = \lambda_3 \sum_{i=1}^{N} \sum_{k=1}^{K} \left(v_{ik}^2 \left\| \mathbf{h}_i^{(M)} - \mathbf{c}_k \right\|_\sigma + P_e(v_{ik}) \right) \tag{17}$$

, where $P_e(v_{ik}) = v_{ik}^2 \ln v_{ik}^2 - v_{ik}^2$, λ_3 acts as a tunable balancing coefficient, ensuring that the values of pertinent data v_{ik} stay elevated, thus preventing the derivation of non-discriminatory, trivial solutions. The enhanced DADPc model now exhibits a monotonic decrease as v_{ik} diminishes. This model utilizes the fuzzy entropy term [i.e., $P_e(v_{ik})$] in Equation 17 to mitigate the adverse effects of noisy data on classification outcomes. The greater the fuzzy entropy is, the more the discriminative information amount of the samples increases, which plays a positive role in enhancing the robust effectiveness of the distribution distance metric. Moreover, fuzzy entropy can effectively limit the influence of noisy or abnormal data in domain distribution alignment. For a comprehensive discussion and empirical insights into how fuzzy entropy enhances robustness, see the analysis in reference (Gretton et al., 2009).

Since each batch of training data includes both the source and target domains in DPC, and since c_k is initialized solely with data from the source domain, it cannot update c_k in real-time during the training stage. To address this issue, we employ a memory bank strategy. The Memory Bank is designed to preserve cluster centroids and their corresponding feature vectors from both domains, which are mapped according to their respective data clusters. We apply the L_2 -norm technique to normalize the feature vectors $h_i^{(M)}$, resulting in normalized features denoted as $\left\| h_i^{(M)} \right\|_2$ alongside the cluster centroids. These values are updated through an iterative process. To estimate real-time probabilities for generating pseudo-labels, the memory bank stores the cluster centroids c_k. Additionally, the feature vectors $h_i^{(M)}$ stored in the memory bank are utilized to compute the latest cluster centroids and update the outdated c_k in the memory bank after each training epoch at the final decoder layer. The initial cluster centroids stored in the memory bank originate from the source domain.

Let $\boldsymbol{B} \in R^{(N+K)\times d}$ be a memory bank that retains the features of all data from both the source and target domains, along with the cluster centroids. Here, d signifies the dimensionality of the features in the final linear layer.

$$\boldsymbol{B} = \left[\boldsymbol{h}_1^{(M)}, \boldsymbol{h}_2^{(M)}, \ldots \boldsymbol{h}_N^{(M)}, \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K \right], \tag{18}$$

where $h_i^{(M)}$ and c undergo L_2 normalization. To account for samples not present in the current mini-batch, we utilize a memory bank to store features and compute similarities, following the approach outlined in Saito et al. (2020). During each iteration, the memory bank $\boldsymbol{B}$ is updated with features from the mini-batch. Let $\boldsymbol{h}_i^{(M)}$ represent the features within the mini-batch, and let T_b denote the set of indices corresponding to the samples in the mini-batch. For every i in T_b, we establish:

$$B_i = h_i^{(M)}. \tag{19}$$

As a result, the memory bank $\boldsymbol{B}$ comprises the recently updated features from the current mini-batch, the older features that are not included in the mini-batch, and the K cluster centroids. Unlike Saito et al. (2020), our approach to updating the memory involves storing features directly, without considering the momentum of features from prior epochs.

4.5 Final formulation

The DADPc model is proposed by embedding the objective function of possibilistic clustering with entropy regularization and adaptive loss defined in Equation 20 using an auto-encoder network as

$$\begin{aligned} &\Theta(\mathbf{W}^{(m)}, \mathbf{b}^{(m)}, \mathbf{V}, \mathbf{C}) = \min_{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}, \mathbf{V}, \mathbf{C}} \left\| \mathbf{H}^{(M)} - \mathbf{X} \right\|_\sigma^2 \\ &- \sum_{i=1}^{n_s} \sum_{k=1}^{K} q_k \log \tilde{p}_k(x_i^s) + \lambda_1 \sum_{m=1}^{M} \left\| \mathbf{W}^{(m)} \right\|_\sigma^2 \\ &+ \lambda_2 \sum_{m=1}^{M} \left\| \mathbf{b}^{(m)} \right\|_{\sigma_1}^2 + \lambda_3 \sum_{i=1}^{N} \sum_{k=1}^{K} \left(v_{ik}^2 \left\| \mathbf{h}_i^{(M)} - \mathbf{c}_k \right\|_\sigma \right. \\ &\left. + v_{ik}^2 \ln v_{ik}^2 - v_{ik}^2 \right) \\ &\text{s.t. } 0 \le v_{ik} \le 1. \end{aligned} \tag{20}$$

Accordingly, DADPc aims to project the raw data onto a nonlinear, low-dimensional feature space and to learn a soft clustering membership matrix using the features mapped nonlinearly simultaneously.

5 Optimization algorithms

In this section, we initially devise an effective algorithm to address the adaptive loss function $\|.\|_\sigma$, which ensures convergence to a local minimum. Subsequently, we introduce an algorithm aimed at solving the loss function associated with DADPc, as outlined in Equation 20.

5.1 Optimization for Weighted Adaptive Loss Function

First, we will examine a broader problem of adaptive loss minimization, formulated as follows:

$$\min_x f(x) + \sum_i \left\| g_i(x) \right\|_\sigma, \tag{21}$$

where $g_i(x)$ yields either a vector or a matrix as output. It is evident that problems $\|.\|_\sigma$ in Equation 20 constitute specific instances of

Equation 21. The Equation 21 can be rewritten as

$$\min_x f(x) + \sum_i \frac{(1+\sigma)\left\|\boldsymbol{g}_i(\mathbf{x})\right\|_2^2}{\left\|\boldsymbol{g}_i(\mathbf{x})\right\|_2 + \sigma}. \tag{22}$$

Building upon the earlier sparse learning optimization algorithms Nie et al. (2013), we introduce an iterative re-weighting approach to address Equation 22. Making the derivative of Equation 22 w.r.t x and equating it to zero, we are able to obtain

$$\nabla f(\mathbf{x}) + 2(1+\sigma)\sum_i \frac{\left\|\boldsymbol{g}_i(\mathbf{x})\right\|_2 + 2\sigma}{2\left(\left\|\boldsymbol{g}_i(\mathbf{x})\right\|_2 + \sigma\right)^2}\nabla \boldsymbol{g}_i(\mathbf{x}) \cdot \boldsymbol{g}_i(\mathbf{x}) = 0, \tag{23}$$

Define

$$d_i = (1+\sigma)\frac{\left\|\boldsymbol{g}_i(\mathbf{x})\right\|_2 + 2\sigma}{2\left(\left\|\boldsymbol{g}_i(\mathbf{x})\right\|_2 + \sigma\right)^2}, \tag{24}$$

Then, Equation 23 can be rewritten as follows:

$$\nabla f(\mathbf{x}) + 2\sum_i d_i \nabla \boldsymbol{g}_i(\mathbf{x}) \cdot \boldsymbol{g}_i(\mathbf{x}) = 0. \tag{25}$$

Note that Equation 25 is still difficult to solve. However, if we fix d_i, then Equation 21 is equivalent to

$$\min_x f(x) + \sum_i d_i \left\|\boldsymbol{g}_i(\mathbf{x})\right\|_2^2, \tag{26}$$

In the given context, the iterative update rule for d_i is defined as $d_i \leftarrow (1+\sigma)\frac{\|\boldsymbol{g}_i(\mathbf{x})\|_2 + 2\sigma}{2(\|\boldsymbol{g}_i(\mathbf{x})\|_2 + \sigma)^2}$. To address Equation 21, we introduce Algorithm 1. The iterative optimization of the adaptive loss function has been demonstrated to converge in Nie et al. (2013). As Algorithm 1 employs an alternating approach to optimize the adaptive loss, the objective function value decreases monotonically, ensuring the convergence of Algorithm 1.

Input: Data vector x.
Output: The current x.

while *Not Converge* **do**
  1. Calculate $d_i = (1+\sigma)\sum_i \frac{\|g_i(x)\|_2 + 2\sigma}{2(\|g_i(x)\|_2 + \sigma)^2}$.
  2. Update x by solving $\min_x f(x) + \sum_i d_i \left\|g_i(x)\right\|_2^2$.
end

Algorithm 1. Algorithm to solve Equation 21.

5.2 Optimization for DADPc

In this subsection, we will present the details regarding the optimization of Equation 20 using an iterative method known as stochastic gradient descent (SGD). For simplicity, we rewrite Equation 20 as

$$\begin{aligned}
&\min_{\Gamma} \mathcal{J} = \\
&\min_{\Gamma} \tfrac{1}{2}\left\|\mathbf{h}_{\mathbf{i}}^{(M)} - \mathbf{x}_{\mathbf{i}}\right\|_{\sigma}^2 - \textstyle\sum_{i=1}^{n_s}\sum_{k=1}^{K} q_k \log \tilde{p}_k(x_i^s) \\
&+\tfrac{\lambda_1}{2}\textstyle\sum_{m=1}^{M}\left\|\mathbf{W}^{(m)}\right\|_{\sigma}^2 \\
&+\tfrac{\lambda_2}{2}\textstyle\sum_{m=1}^{M}\left\|\mathbf{b}^{(m)}\right\|_{\sigma_1}^2 + \lambda_3\textstyle\sum_{i=1}^{N}\sum_{k=1}^{K}\left(v_{ik}^2\left\|\mathbf{h}_i^{(M)} - \mathbf{c}_k\right\|_{\sigma}\right. \\
&\left.+v_{ik}^2 \ln v_{ik}^2 - v_{ik}^2\right) \\
&\text{s.t. } 0 \le v_{ik} \le 1
\end{aligned} \tag{27}$$

where $\mathbf{\Gamma} = \left\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}, \mathbf{V}, \mathbf{C}\right\}$ in order to keep the equations unclustered. As shown in Subsection 5.1, Equation 27 is equivalent to the following dual:

$$\begin{aligned}
&\min_{\Gamma} \mathcal{L} = \\
&\min_{\Gamma} \tfrac{1}{2}a_{ik}\left\|\mathbf{h}_{\mathbf{i}}^{(M)} - \mathbf{x}_{\mathbf{i}}\right\|_2^2 - \textstyle\sum_{i=1}^{n_s}\sum_{k=1}^{K} q_k \log \tilde{p}_k(x_i^s) \\
&+\tfrac{\lambda_1}{2}\textstyle\sum_{m=1}^{M} r_{ik}^w\left\|\mathbf{W}^{(m)}\right\|_{\sigma}^2 + \tfrac{\lambda_2}{2}\textstyle\sum_{m=1}^{M} r_{ik}^b\left\|\mathbf{b}^{(m)}\right\|_{\sigma_1}^2 \\
&+\lambda_3\textstyle\sum_{i=1}^{N}\sum_{k=1}^{K}\left(e_{ik}v_{ik}^2\left\|\mathbf{h}_i^{(M)} - \mathbf{c}_k\right\|_2 + v_{ik}^2 \ln v_{ik}^2 - v_{ik}^2\right) \\
&\text{s.t. } 0 \le v_{ik} \le 1
\end{aligned} \tag{28}$$

where

$$a_{ik} = (1+\sigma)\frac{\left\|\mathbf{h}_i^{(M)} - \mathbf{x}_i\right\|_2 + 2\sigma}{2\left(\left\|\mathbf{h}_i^{(M)} - \mathbf{c}_k\right\|_2 + \sigma\right)^2} \tag{29}$$

$$r_{ik}^w = (1+\sigma)\frac{\left\|\mathbf{W}^{(m)}\right\|_2 + 2\sigma}{2\left(\left\|\mathbf{W}^{(m)}\right\|_2 + \sigma\right)^2} \tag{30}$$

$$r_{ik}^b = (1+\sigma_1)\frac{\left\|\mathbf{b}^{(m)}\right\|_2 + 2\sigma_1}{2\left(\left\|\mathbf{b}^{(m)}\right\|_2 + \sigma_1\right)^2} \tag{31}$$

$$e_{ik} = (1+\sigma)\frac{\left\|\mathbf{h}_i^{(M)} - \mathbf{c}_k\right\|_2 + 2\sigma}{2\left(\left\|\mathbf{h}_i^{(M)} - \mathbf{c}_k\right\|_2 + \sigma\right)^2} \tag{32}$$

We can solve Equation 27 by utilizing the coordinate blocking method.

Update $\mathbf{W}^{(m)}, \mathbf{b}^{(m)}$ by fixing $\mathbf{V}, \mathbf{C}$: According to the definition of $\mathbf{h}_i^{(m)}$ and the back-propagation algorithm, the subgradient of Equation 28 w.r.t. $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ can be derived as

$$\begin{cases} \nabla_{\mathbf{W}^{(m)}}\mathcal{L} = \mathbf{\Delta}_i^{(m)}\mathbf{h}_i^{(m-1)T} + \lambda_1 r_{ik}^w \mathbf{W}^{(m)} \\ \nabla_{\mathbf{b}^{(m)}}\mathcal{L} = \mathbf{\Delta}_i^{(m)} + \lambda_2 r_{ik}^b \mathbf{b}^{(m)} \end{cases} \tag{33}$$

where $\Delta^{(m)}$ is defined as follows:

$$\mathbf{\Delta}_i^{(m)} = \begin{cases} \left(\mathbf{W}^{(m+1)}\mathbf{\Delta}_i^{(m+1)}\right) \odot f'\left(\mathbf{a}_i^{(m)}\right), m \ne M \\ \left(\mathbf{\Theta}_{1i}^{(M)} - \lambda_1\mathbf{\Theta}_{2i}^{(M)} + \lambda_3\mathbf{\Theta}_{3i}^{(M)}\right) \odot f'\left(\mathbf{a}_i^{(M)}\right), m = M \end{cases} \tag{34}$$

where $\Theta_{1i}^{(M)}$ and $\Theta_{2i}^{(M)}$ are defined as

$$\mathbf{\Theta}_{1i}^{(M)} = \left(\mathbf{h}_i^{(M)} - h_i^{(0)}\right)a_{ik}, \tag{35}$$

$$\mathbf{\Theta}_{2i}^{(M)} = \sum_{i=1}^{n_s} \sum_{k=1}^{K} q_k \left(1 - \frac{\exp\left(h_{ik}^{(M)}\right)}{\sum_{j=1}^{K} \exp\left(h_{ij}^{(M)}\right)}\right), \tag{36}$$

$$\mathbf{\Theta}_{3i}^{(M)} = \left(\mathbf{H}^{(M)} - \mathbf{Y}_{MB}^{(M)} \mathbf{C}_{MB}^{(M)}\right) \mathbf{e}_{ik} \mathbf{v}_{ik}^2 \tag{37}$$

$\odot$ is the element-wise multiplication, $f'(\cdot)$ is derivative of the activation function $f(\cdot)$, and $\mathbf{a}_i^{(m)}$ is the input of m-th layer, i.e., $\mathbf{a}_i^{(m)} = \mathbf{W}^{(m)} \mathbf{h}_i^{(m-1)} + \mathbf{b}^{(m)}$. $\mathbf{Y}_B^{(M)} \in \mathbb{R}^{N \times K}$. $\mathbf{C}_B^{(M)} \in \mathbb{R}^{K \times d}$ is from memory bank. According to the cluster centroid $c_k \in \mathbf{C}_B^{(M)}$, we can obtain the label of each feature vector $y_i \in \mathbf{Y}_B^{(M)}$ in the M-th layer. $\mathbf{Y}_B^{(M)} \mathbf{C}_B^{(M)} \in \mathbb{R}^{N \times d}$. Each row in $\mathbf{Y}_B^{(M)} \mathbf{C}_B^{(M)}$ corresponds to the cluster centroid of each feature vector.

Based on Equations 33–37 using the SGD algorithm, we update $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ as follows:

$$\begin{cases} \mathbf{W}^{(m)} = \mathbf{W}^{(m)} - \mu \nabla_{\mathbf{W}^{(m)}} \mathcal{F} \\ \mathbf{b}^{(m)} = \mathbf{b}^{(m)} - \mu \nabla_{\mathbf{b}^{(m)}} \mathcal{F} \end{cases} \tag{38}$$

Using the initial K cluster centroids from the source domain, we then acquire new pseudo-labels for the target domain by employing a nearest-centroid classification approach: $y_t = \arg\min_k D\left(h_{(t,j)}^{(M)}, c_k\right)$. The distance between a and b is quantified by $D(a, b)$. By default, our chosen measure of distance is the cosine similarity metric. Since the feature vectors from $h_i^{(M)}$ and the initial K cluster centroids are retained in the memory bank, we update $\mathbf{C}$ and y_t according to Equation 39 (i.e., $\mathbf{C}_B^{(M)}$) after each batch of data training:

$$c_k = \frac{\sum_{i=1}^{N} 1\left(\hat{y}_t = k\right) h_i^{(M)}}{\sum_{\hat{y} \in Y_s, \hat{Y}_t} 1\left(\hat{y} = k\right)} \tag{39}$$

Updating $\mathbf{V}$ by fixing $\mathbf{W}^{(m)}$, $\mathbf{b}^{(m)}$ and $\mathbf{C}$ through optimizing Equation 20 directly: When $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ are fixed, Equation 20 becomes equivalent to:

$$\begin{aligned} &min_{v_{ik}} \sum_{i=1}^{N} \sum_{k=1}^{K} \lambda_3 \left(v_{ik}^2 \left\|\mathbf{h}_i^{(M)} - \mathbf{c}_k\right\|_\sigma + v_{ik}^2 \ln v_{ik}^2 - v_{ik}^2\right), \\ &\text{s.t. } 0 \le v_{ik} \le 1 \end{aligned} \tag{40}$$

The Lagrangian function for Equation 40 is represented as

$$\begin{aligned} &min_{v_{ik}} \lambda_3 \left(\sum_{i=1}^{N} \sum_{j=1}^{k} v_{ik}^2 \left\|\mathbf{h}_i^{(M)} - \mathbf{c}_k\right\|_\sigma + v_{ik}^2 \ln v_{ik}^2 - v_{ik}^2\right) \\ &- \sum_{i=1}^{N} \sum_{j=1}^{k} \beta_{ik} \left(v_{ik}\right) \end{aligned} \tag{41}$$

where β_{ik} is aa Lagrangian multiplier with $0 < \beta_{ik} < 1$. According to KKT conditions, we have

$$v_{ik} = \exp\left(\frac{1 - \lambda_3 \sum_{i=1}^{N} \sum_{k=1}^{K} \left\|\mathbf{h}_i^{(M)} - \mathbf{c}_k\right\|_\sigma}{4}\right) \tag{42}$$

We can therefore apply an iterative algorithm to update $\mathbf{W}^{(m)}$, $\mathbf{b}^{(m)}$, $\mathbf{V}$ and $\mathbf{C}$. If SGD decreases $\mathcal{L}$ in Equation 28, the algorithm will converge to a local minimum, as the optimization can be viewed as a variant of coordinate gradient descent.

The optimization procedure of Equation 27 is summarized in Algorithm 2.

Input: Labeled data: $\{X_s, Y_s\} = \{(\mathbf{x}_1^s, y_1^s), \ldots, (\mathbf{x}_{n_s}^s, y_{n_s}^s)\}$;
Unlabeled data: $X_t = \{\mathbf{x}_1^t, \ldots, \mathbf{x}_{n_t}^t\}$;
$\mathbf{V}_0$ is initialized by Equation 42, the number of clusters k, parameters λ_1, λ_2, λ_3, β, SGD maximum iterations L.
Initialize: $\mathbf{H}^{(0)} = \mathbf{X}$, $\mathbf{V} = \mathbf{V}_0$, a random matrix $\mathbf{C} \in \mathbb{R}^{d^{(M)} \times k}$, random matrices $\mathbf{W}^{(m)}$ and random vectors $\mathbf{b}^{(m)}$ where $m = 1, 2, \cdots, M$.
Pre-train the auto-encoder.
Output: clustering assignment matrix $\mathbf{V}$, centroid matrix $\mathbf{C}$ and parametric neural network with $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$.

/* The iterative steps empirically converge in fewer than 40 iterations, as illustrated in Subsection 6.4.3. */
while $\mathbf{C}$ *does not converge or* $\mathbf{V}$ *doesn't exceed the maximum iterations* **do**
  Calculate a_{ik}, r_{ik}^w, r_{ik}^b, e_{ik} by Equations 29, 30, 31, 32 for $i = 1, 2, \cdots, N$ and $k = 1, 2, \cdots, K$.
  for *iter* $= 1$ *to* L **do**
    Update $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ by Equation 38 for $m = 1, 2, \cdots, M$.
  end
  Update C **by** Equation 39.
  Update V **by** Equation 42.
end

Algorithm 2. Algorithm to solve Equation 27.

6 Experimental analysis

This section employs three well-known benchmark datasets: SEED (Zheng and Lu, 2015), SEED-IV (Zheng and Lu, 2013; Zhou et al., 2023), and DEAP (Sander et al., 2012), to thoroughly assess the model's ability to recognize EEG-based emotions. Specifically, in the following tables of experimental results, the bold values in each table are the best accuracy performance results achieved. Pacc denotes the average accuracy performance of each method.

6.1 Datasets

6.1.1 Description

The SEED dataset includes EEG data on emotions collected from 15 individuals who watched 15 movie clips designed to elicit three emotional responses: positive, negative, and neutral. Conversely, the SEED-IV dataset contains EEG emotion data from 15 participants as well but features 24 movie segments aimed at evoking four emotions: happiness, sadness, neutrality, and fear. Both the SEED and SEED-IV datasets involved participants engaging in three sessions, each occurring on separate days with a one-week interval between them. The EEG signals were recorded using a 62-channel ESI Neuroscan system.

Moreover, we are keen to investigate performance in cross-dataset scenarios, specifically exploring whether acceptable recognition precision can be sustained when the training and testing data originate from different subjects, are recorded with varying EEG equipment, and involve emotional states elicited by diverse stimuli. Furthermore, our objective is to assess if multi-source domain adaptation methods can enhance results in such circumstances. To achieve this, we utilize the DEAP dataset (Sander et al., 2012), which is publicly available, to examine emotional states. This dataset includes data from 32 individuals who viewed 40 one-minute music videos intended to evoke emotional reactions, while their physiological responses were recorded. After watching each video, the participants evaluated their emotions in five aspects: valence (how pleasant it was), arousal (the level of excitement), dominance (degree of control), liking (personal preference), and familiarity (recognition of the stimulus). The evaluation scores range from one (the lowest level) to nine (the highest level), but for familiarity, the scores vary from one to five.

To gain a deeper understanding of these three benchmarks, please refer to Lan et al. (2019). As stated in Zhong et al. (2022) and Lan et al. (2019), there are notable differences among these benchmarks. These discrepancies can arise from various factors, including variations in sessions, participants, experimental procedures, EEG equipment, and the types of emotional stimuli used.

6.1.2 Feature extraction

The SEED dataset's EEG data preprocessing followed standard procedures outlined in Luo et al. (2024). First, the EEG signals were downsampled to 200 Hz. Next, artifacts such as electrooculogram (EOG) and electromyography (EMG) were removed. A bandpass filter ranging from 0.3 to 50 Hz was applied to enhance signal quality. Each trial was then divided into several 1-second data segments. The length of each trial in the SEED dataset varied between 185 and 265 seconds, depending on the duration of the emotional stimulus used to evoke the targeted emotion. To ensure consistency across various classes, all trials were shortened to a uniform length of 185s.

The DEAP dataset's EEG data were recorded using Biosemi Active Two devices, initially sampled at 512 Hz and subsequently down-sampled to 128 Hz. In DEAP, emotions are scored on a five-point scale; to align with the SEED dataset, we discretize the emotional dimensions as follows: positive for valence ratings above 7, neutral for ratings between 3 and 7, and negative for ratings below 3. We identify DEAP trials where the majority of participants reported the successful induction of positive, neutral, and negative emotions. Specifically, trial 18 evokes positive emotion, trial 16 evokes neutral emotion, and trial 38 evokes negative emotion, with 27, 28, and 19 participants, respectively, confirming the intended emotion. The participants who consistently reported successful emotion elicitation in these trials (18, 16, and 38) are subjects 2, 5, 10, 11, 12, 13, 14, 15, 19, 22, 24, 26, 28, and 31. Therefore, for the DEAP dataset, we only use the selected trials from these fourteen subjects. Each trial lasts 63 seconds, but the first 3 seconds consist of baseline recordings without emotion elicitation; thus, we use the segment starting from the 4th second, resulting in a 60-second valid trial length.

To capture emotion-related information, we computed differential entropy (DE) features (Zhong et al., 2022; Lan et al., 2019) for five frequency bands: Delta (1–3 Hz), Theta (4–7 Hz), Alpha (8–13 Hz), Beta (14–30 Hz), and Gamma (31–50 Hz). This process generated 310 features (62 channels $\times$ 5 frequency bands) for every 1-second data segment, serving as model input. For the DEAP dataset, the ultimate feature vector comprised 160 dimensions (32 channels $\times$ 5 frequency bands), with each trial containing 60 samples. In the SEED dataset, the final feature vector had 310 dimensions (62 channels $\times$ 5 frequency bands), with each trial yielding 185 samples.

Following the approach in Donahue et al. (2014), our DADPc method can be readily trained utilizing deep features extracted from established models. We fine-tune pre-trained deep models, such as Resnet101 (EEG differential entropy features (62 channels $\times$ 5 bands) are reshaped into 62×5 matrices, mimicking image input for ResNet-101). The Resnet101's bottleneck layer outputs are used to balance semantic and spatial information for cross-domain alignment. He et al. (2016), on the source domain and then extract deep features from EEG signals.The Resnet101's Bottleneck layer outputs are used to balance semantic and spatial information for cross-domain alignment. Pre-experiments showed ResNet-101's deep residual structure outperforms ResNet-50/DenseNet in cross-domain scenarios, with mature open-source implementations aiding reproducibility.. These deep representations can subsequently be used to train the recognition model.

6.2 Experimental protocols

Since the parameter $\mathbf{b}$ does not require feature selection, the $l_{\sigma 1}$-norm on $\mathbf{b}$ can be fixed as the l_2-norm. To thoroughly assess the robustness and consistency of our proposed DADPc method and facilitate comparison with previous works, we utilize four unique validation protocols that incorporate various evaluation strategies for a comprehensive examination, as outlined in Zhou et al. (2022); Tao et al. (2024):

1) Cross-subject cross-session (CUCE) with leave-one-subject-out (LOSO) cross-validation. To rigorously evaluate the model's durability on novel subjects and sessions, we apply a strict assessment protocol known as CUCE with LOSO. During each cycle, the session data of one subject is designated as the target, while the session data from all other subjects are utilized as the source. This training and validation procedure is repeated until each subject's sessions have served as the target once. Given the inherent variability among individuals and sessions, this evaluation protocol poses a significant challenge for the model's proficiency in EEG-based emotion recognition tasks.

2) Cross-subject single-session (CUSE) with LOSO cross-validation. The most commonly adopted validation technique in EEG-based emotion recognition (Li et al., 2019a; Luo et al., 2018; Li et al., 2019b; Zhou et al., 2022; Luo et al., 2024) involves assigning data from a single session of one subject as the target, while data from all other subjects serve as the source. This iterative training and validation process continues until each subject has been designated as the target once. Consistent with other research, this cross-validation method considers only the first session.
3) Within-subject Cross-session (WUCE) with LOSO cross-validation. Consistent with prevalent techniques, we use a time series cross-validation methodology that leverages historical data to forecast present or future data. In this context, the first two sessions for each subject serve as the source domain, while the following session acts as the target domain. The final results are obtained by calculating the mean accuracy and standard deviation across all subjects.
4) Cross-database cross-validation (CDCV). In line with the configurations specified in Tao and Dan (2021); Tao et al. (2023), we employ 32 common channels from SEED and DEAP to construct a unified feature space of 160 dimensions, enhancing cross-dataset generalization. This enables the formulation of multiple cross-dataset generalization scenarios: DEAP → SI, DEAP → SII, DEAP → SIII, SI → DEAP, SII → DEAP, and SIII → DEAP. Here, "A→ B" denotes adapting from dataset A to dataset B, with SI, SII, and SIII denoting Session I, Session II, and Session III of SEED, respectively. When DEAP is the source dataset, 2,520 data points are selected, whereas 2775 data points are chosen from each of the three SEED sessions (SI, SII, and SIII) to serve as the target datasets. Conversely, when each SEED session is the source, we resample 41,625 data points for training and 180 samples from DEAP as the target.

6.3 Experimental results

Given that parameter selection remains a pressing issue in machine learning, this study empirically investigates the parameter space to identify the most effective settings. Our approach involves three primary model parameters, initially set to λ_1, λ_2, and λ_3 equal to 0. These parameters are then fine-tuned using cross-validation within the range $\left[10^{-4}, 10^{5}\right]$. Moreover, for the experimental setup, the matrix norm is set at $\sigma = 0.001$. The results are reported as the average performance across all participants.

To compare our method with several recently introduced deep adaptation models, we specifically assess DADPc using deeply extracted features from the neural network with five layers (i.e., 1024-500-300-500-1024). For other deep benchmarks, we directly utilize their publicly available source codes to fine-tune the pre-trained models employed in their respective works.

6.3.1 CUCE with LOSO cross-validation

To assess the effectiveness and consistency of our model across both cross-subject and cross-session contexts, we employ LOSO cross-validation to validate the proposed DADPc approach within the CUCE framework, utilizing the SEED and SEED-IV datasets. The results, presented in Tables 1, 2, reveal that our method outperforms others on both datasets, achieving an emotion recognition accuracy of 87.83 ± 4.21% for three classes on SEED and 75.31 ± 6.22% for four classes on SEED-IV. These findings underscore the superior recognition precision and enhanced generalization capability of the DADPc method, particularly in the presence of complex individual and environmental variations, which suggests improved emotional validity.

6.3.2 CUSE with LOSO cross-validation

Tables 3, 4 summarize the experimental outcomes of the LOSO recognition task conducted on the SEED and SEED-IV datasets within the CUSE framework, alongside a benchmarking against previous studies. All results are reported as mean ± standard deviation. Our proposed DADPc model, as shown in these tables, achieves a peak performance of 94.62% with a standard deviation of 4.37%. DADPc outperforms the best-reported literature results by 1.56%, exhibiting a lower standard deviation on the SEED-IV dataset. Notably, the model's performance is superior on the SEED-IV dataset compared to the SEED dataset. This highlights DADPc's efficacy in addressing individual differences and improving robust pseudo-labeling (Litrico et al., 2023) for a wider range of emotion recognition in affective Brain-Computer Interface (aBCI) applications.

6.3.3 WUCE with LOSO cross-validation

The outcomes of the WUCE cross-validation for the SEED dataset are outlined in Table 5, and those for the SEED-IV dataset are presented in Table 6.

During the experiments on the SEED datasets, EEGMatch emerged as the top-performing method. This can be attributed to the mixup technique, which enriched the data and enhanced model training, albeit at the cost of higher computational expenses due to the increased data volume. Nonetheless, DADPc achieved results that were comparable, closely trailing behind. This underscores DADPc's proficiency in categorizing distinct classes. For the SEED-IV dataset, which involves four-class emotion recognition, DADPc excelled, particularly as the number of categories increased. This highlights DADPc's superior accuracy in recognizing more nuanced emotions and its robust scalability.

6.3.4 CDCV results

In this section, we aim to evaluate the extensive and consistent generalization capability of our proposed DADPc method, particularly in the realm of cross-dataset emotion recognition. Essentially, achieving generalization across datasets presents a more formidable challenge than cross-subject generalization due to the significant differences that exist between datasets.

The experimental outcomes for six tasks, as shown in Table 7, indicate that the performance of all methods when applied across datasets is slightly lower compared to their performance within the same dataset. This finding supports the notion that

TABLE 1 The mean accuracies (%) and standard deviations (%) on the SEED database using CUCE with LOSO cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
RF; Breiman (2001)	69.60 ± 7.64	KNN (Duda et al., 2000)	60.66 ± 7.93
SVM; Vapnik (1995)	62.24 ± 5.48	Adaboost (Zhu et al., 2006)	71.87 ± 5.70
TCA; Pan et al. (2011)	65.31 ± 6.04	CORAL (Sun et al., 2015)	69.22 ± 4.11
SA; Fernando et al. (2013)	61.41 ± 9.75	GFK (Gong et al., 2012)	67.36 ± 6.52
DICE; Liang et al. (2019)	73.56 ± 4.23	GAKT (Ding et al., 2018)	74.82 ± 7.14
MDDD; Luo et al. (2024)	76.60 ± 6.79	EDPC	**76.82 ± 6.14**
Deep learning methods			
DCORAL; Sun and Saenko (2016)	80.87 ± 6.04	DAN (Long et al., 2019)	82.51 ± 3.71
DDC; Tzeng et al. (2014)	82.17 ± 4.96	DANN (Ganin et al., 2016)	84.79 ± 6.44
PR-PL; Zhou et al. (2022)	85.56 ± 4.78	PARSE (Zhang and Etemad, 2022)	82.44 ± 5.00
EEGMatch; Zhou et al. (2023)	86.30 ± 5.04	DADPc	**87.83 ± 4.21**

TABLE 2 The mean accuracies (%) and standard deviations (%) on the SEED-IV database using CUCE with LOSO cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
KNN; Duda et al. (2000)	40.83 ± 7.28	SVM (Vapnik, 1995)	51.78 ± 12.85
Adaboost; Zhu et al. (2006)	53.44 ± 9.12	TCA (Pan et al., 2011)	56.56 ± 13.77
CORAL; Sun et al. (2015)	49.44 ± 9.09	SA (Fernando et al., 2013)	64.44 ± 9.46
GFK; Gong et al. (2012)	45.89 ± 8.27	KPCA (Schlkopf et al., 1998)	51.76 ± 12.89
DICE; Liang et al. (2019)	66.75 ± 7.25	GAKT (Ding et al., 2018)	64.48 ± 5.52
MDDD; Luo et al. (2024)	64.90 ± 10.25	EDPC	**67.88 ± 5.21**
Deep learning methods			
DGCNN; Song et al. (2018)	52.82 ± 9.23	DAN (Long et al., 2019)	58.87 ± 8.13
RGNN; Zhong et al. (2020)	73.84 ± 8.02	BiHDM (Li et al., 2019c)	69.03 ± 8.66
BiDANN; Li et al. (2018b)	65.59 ± 10.39	DANN (Ganin et al., 2016)	54.63 ± 8.03
PR-PL; Zhou et al. (2022)	74.92 ± 7.92	PARSE (Zhang and Etemad, 2022)	69.78 ± 8.22
EEGMatch; Zhou et al. (2023)	73.60 ± 7.53	DADPc	**75.31 ± 6.22**

the distributional differences between two datasets are more pronounced than those between two subjects.

Specifically, our DADPc model demonstrates superior performance compared to other baseline methods in 4 out of the 6 recognition tasks. While CAN (Kang et al., 2022) occasionally achieves the best results in two particular settings, DADPc consistently ranks first in other situations. These results suggest that the combined approach of reconstructing feature learning and possibilistic clustering learning is a more effective strategy.

Finally, as observed from Tables 1–7, the performance of DADPc consistently outperforms that of EDPC, achieving a maximum improvement of nearly 18%. This result suggests that utilizing a simple deep neural network for feature extraction and reconstruction can effectively enhance emotion recognition performance.

6.4 Discussion

To thoroughly assess the model's efficacy, we conduct additional evaluations to determine the impact of various configurations within the DADPc framework.

6.4.1 Effect of noisy labels

To evaluate the model's resilience in scenarios with noisy labels, we randomly introduce η% noise to the source labels and assess the model's performance on unseen target data. In particular, we replace η% of the actual labels in Y_s with random labels and then conduct supervised learning. Afterward, we test the trained model on the target domain. It is important to note that noise is only introduced to the source domain, while the target domain is used

TABLE 3 The mean accuracies (%) and standard deviations (%) on the SEED database using CUSE with LOSO cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
TKL; Long et al. (2015)	63.54 ± 15.47	T-SVM (Ronan et al., 2006)	68.57 ± 9.54
TCA; Pan et al. (2011)	63.64 ± 14.88	TPT (Zheng and Lu, 2013)	73.86 ± 11.05
KPCA; Schlkopf et al. (1998)	61.28 ± 14.62	GFK (Gong et al., 2012)	71.31 ± 14.09
SA; Fernando et al. (2013)	66.00 ± 10.89	DICA (Ma et al., 2019)	69.40 ± 07.80
DBN; Zheng and Lu (2015)	61.01 ± 12.38	SVM (Vapnik, 1995)	58.18 ± 13.85
DICE; Liang et al. (2019)	74.22 ± 7.33	GAKT (Ding et al., 2018)	72.29 ± 4.66
MDDD; Luo et al. (2024)	**84.57 ± 9.49**	EDPC	82.34 ± 4.52
Deep learning methods			
DGCNN; Song et al. (2018)	79.95 ± 9.02	DAN (Long et al., 2019)	83.81 ± 8.56
RGNN; Zhong et al. (2020)	85.30 ± 6.72	BiHDM (Li et al., 2019c)	85.40 ± 7.53
WGAN-GP; Luo et al. (2018)	87.10 ± 7.10	MMD (Dino et al., 2013)	80.88 ± 10.10
ATDD-DANN; Du et al. (2022)	90.92 ± 1.05	JDA-Net (Li et al., 2019b)	88.28 ± 11.44
R2G-STNN; Li et al. (2022)	84.16 ± 7.63	SimNet (Pinheiro, 2018)	81.58 ± 5.11
BiDANN; Li et al. (2018b)	83.28 ± 9.60	DResNet (Ma et al., 2019)	85.30 ± 8.00
ADA; Philip et al. (2017)	84.47 ± 10.65	DANN (Ganin et al., 2016)	81.65 ± 9.92
PR-PL; Zhou et al. (2022)	93.06 ± 5.12	PARSE (Zhang and Etemad, 2022)	82.11 ± 5.83
EEGMatch; Zhou et al. (2023)	92.45 ± 6.85	DADPc	**93.58 ± 6.35**

TABLE 4 The mean accuracies (%) and standard deviations (%) on the SEED-IV database using CUSE with LOSO cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
TKL; Long et al. (2015)	63.54 ± 15.47	T-SVM (Ronan et al., 2006)	68.57 ± 9.54
TCA; Pan et al. (2011)	63.64 ± 14.88	TPT (Zheng and Lu, 2013)	73.86 ± 11.05
KPCA; Schlkopf et al. (1998)	61.28 ± 14.62	GFK (Gong et al., 2012)	71.31 ± 14.09
SA; Fernando et al. (2013)	66.00 ± 10.89	DICA (Ma et al., 2019)	69.40 ± 07.80
DBN; Zheng and Lu (2015)	61.01 ± 12.38	SVM (Vapnik, 1995)	58.18 ± 13.85
DICE; Liang et al. (2019)	74.22 ± 7.33	GAKT (Ding et al., 2018)	72.29 ± 4.66
MDDD; Luo et al. (2024)	76.60 ± 6.79	EDPC	**76.82 ± 7.14**
Deep learning methods			
DGCNN; Song et al. (2018)	79.95 ± 9.02	DAN (Long et al., 2019)	83.81 ± 8.56
RGNN; Zhong et al. (2020)	85.30 ± 6.72	BiHDM (Li et al., 2019c)	85.40 ± 7.53
WGAN-GP; Luo et al. (2018)	87.10 ± 7.10	MMD (Dino et al., 2013)	80.88 ± 10.10
ATDD-DANN; Du et al. (2022)	90.92 ± 1.05	JDA-Net (Li et al., 2019b)	88.28 ± 11.44
R2G-STNN; Li et al. (2022)	84.16 ± 7.63	SimNet (Pinheiro, 2018)	81.58 ± 5.11
BiDANN; Li et al. (2018b)	83.28 ± 9.60	DResNet (Ma et al., 2019)	85.30 ± 8.00
ADA; Philip et al. (2017)	84.47 ± 10.65	DANN (Ganin et al., 2016)	81.65 ± 9.92
PR-PL; Zhou et al. (2022)	93.06 ± 5.12	PARSE (Zhang and Etemad, 2022)	82.11 ± 5.83
EEGMatch; Zhou et al. (2023)	92.45 ± 06.85	DADPc	**94.62 ± 4.37**

TABLE 5 The mean accuracies (%) and standard deviations (%) on the SEED database using WUCE with LOSO cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
RF; Breiman (2001)	76.42 ± 11.15	KNN (Duda et al., 2000)	72.96 ± 12.10
TCA; Pan et al. (2011)	77.63 ± 11.49	CORAL (Sun et al., 2015)	82.18 ± 9.81
SA; ;Fernando et al. (2013)	67.79 ± 7.43	GFK (Gong et al., 2012)	79.28 ± 7.44
DICE; Liang et al. (2019)	81.58 ± 7.55	GAKT (Ding et al., 2018)	80.31 ± 6.44
MDDD; Luo et al. (2024)	81.27 ± 5. 47	EDPC	**82.31 ± 6.44**
Deep learning methods			
DAN; Long et al. (2019)	89.16 ± 7.90	SimNet (Pinheiro, 2018)	86.88 ± 7.83
DDC; Tzeng et al. (2014)	91.14 ± 5.61	ADA (Philip et al., 2017)	89.13 ± 7.13
DANN; Ganin et al. (2016)	89.45 ± 6.74	MMD (Dino et al., 2013)	84.38 ± 12.05
JDA-Net; Li et al. (2019b)	91.17 ± 8.11	DCORAL (Sun and Saenko, 2016)	88.67 ± 6.25
PR-PL; Zhou et al. (2022)	93.18 ± 6.55	PARSE (Zhang and Etemad, 2022)	89.85 ± 5.06
EEGMatch; Zhou et al. (2023)	**94.70 ± 4.10**	DADPc	93.18 ± 5.40

TABLE 6 The mean accuracies (%) and standard deviations (%) on the SEED-IV database using WUCE with LOSO cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
RF; Breiman (2001)	60.27 ± 16.36	KNN (Duda et al., 2000)	54.18 ± 16.28
TCA; Pan et al. (2011)	59.49 ± 12.07	CORAL (Sun et al., 2015)	66.88 ± 14.67
SA; Fernando et al. (2013)	56.94 ± 11.45	GFK (Gong et al., 2012)	60.66 ± 10.00
DICE; Liang et al. (2019)	69.68 ± 12.52	GAKT (Ding et al., 2018)	68.77 ± 6.00
MDDD; Luo et al. (2024)	68.81 ± 9.25	EDPC	**71.39 ± 5.22**
Deep learning methods			
DCORAL; Sun and Saenko (2016)	65.10 ± 13.20	DAN (Long et al., 2019)	60.20 ± 10.20
DDC; Tzeng et al. (2014)	68.80 ± 16.60	MEERNet (Chen et al., 2021)	72.10 ± 14.10
PR-PL; Zhou et al. (2022)	74.62 ± 14.15	PARSE (Zhang and Etemad, 2022)	70.24 ± 8.47
EEGMatch; Zhou et al. (2023)	72.91 ± 8.34	DADPc	**75.36 ± 5.13**

exclusively for evaluating the model. In our experiments, we vary η% at 5%, 10%, 15%, 20%, and 25%. The model accuracies on the SEED dataset for these noise levels are shown in Figure 2, revealing a minor decline in performance as the label noise ratio rises from 5% to 25%. These results demonstrate that DADPc is a robust model with a high level of tolerance for noisy labels. Building on the study in Zhou et al. (2022), our future research could incorporate the recent method (Jin et al., 2023) to further reduce general noise in EEG signals and improve model stability in cross-subject ER applications.

6.4.2 Visualization and confusion matrix

We use the t-distributed stochastic neighbor embedding (t-SNE) algorithm (Laurens and Hinton, 2008) to visually compare the learning ability of our DADPc at various training stages. The visual results are shown in Figure 3. By comparing Figures 3a–c, we observe:

- The intra-class variation across domains in Figure 3a is larger than in Figures 3b, c. This indicates that deep feature extraction and possibilistic clustering enhance DADPc's efficiency.
- In Figure 3c, feature clusters are denser and less scattered than in Figure 3a. In DADPc, each target feature is drawn to its class-cluster center, while target-cluster centers align with source-cluster centers. This demonstrates that DADPc acquires meaningful features and cluster centers, highlighting its advantage in unsupervised domain adaptation.

To qualitatively assess the model's performance across emotion categories, we visually examined the DADPc model confusion matrix on the SEED dataset using the WUCE metric and compared

TABLE 7 The mean accuracies (%) using CDCV (SI, Session I; SII, Session II; SIII, Session III).

Methods	DEAP→SI	DEAP→SII	DEAP→SIII	SI→DEAP	SII→DEAP	SIII→DEAP
Non-Deep learning methods						
SA; Fernando et al. (2013)	56.69	59.33	52.28	55.61	48.90	50.02
TPT; Zheng and Lu (2013)	58.23	60.22	55.39	60.01	51.41	52.23
GAKT; Ding et al. (2018)	60.36	61.33	59.40	59.79	52.49	54.16
MDDD; Luo et al. (2024)	61.29	62.15	61.05	**61.81**	**56.16**	**56.68**
DICE; Liang et al. (2019)	60.68	62.79	60.86	60.49	54.78	55.33
EDPC	**62.17**	**63.36**	**62.08**	60.11	54.89	56.45
Deep learning methods						
DDG; Ding et al. (2018)	62.40	64.92	73.92	64.29	54.29	53.33
DDC; Tzeng et al. (2014)	60.89	62.43	69.43	62.16	52.16	50.07
DANN; Ganin et al. (2016)	61.08	62.51	72.51	63.77	53.77	52.62
DSAN; Zhu et al. (2021)	63.28	64.50	74.50	64.58	55.58	54.10
DCORAL; Sun and Saenko (2016)	60.15	60.42	70.42	61.54	52.54	51.00
CAN; Kang et al. (2022)	64.22	65.77	**75.77**	**66.12**	57.12	55.39
DADPc	**65.83**	**66.30**	75.16	65.39	**57.59**	**58.28**

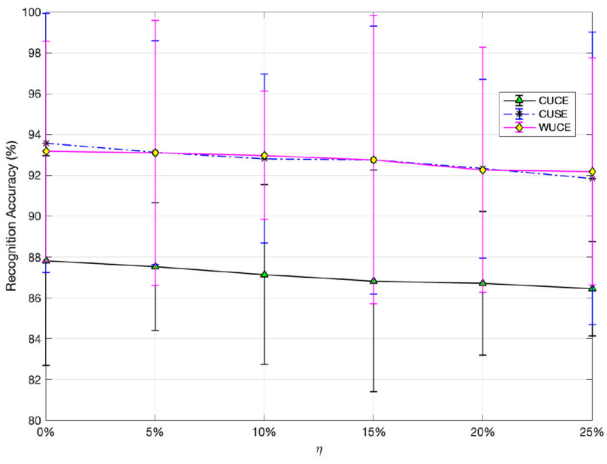

FIGURE 2
Robustness with different ratios of noise labels.

it with recent models (Li et al., 2018b, 2019c; Zhou et al., 2022). Figure 4 shows that all models distinguish positive emotions well (accuracy exceeding 90%) but struggle to differentiate negative and neutral emotions. BiDANN (Li et al., 2019c) has a recognition rate of less than 80% for neutral emotions (76.72%). Compared to existing methods (Figures 4a–c our model demonstrates better recognition, particularly for neutral-negative emotions. As shown in Figure 4d, our model achieves recognition rates of 97.14%, 96.60%, and 97.83% for negative, neutral, and positive emotions, respectively, outperforming PR-PL and highlighting its adaptability and discriminative power in the target domain.

6.4.3 Convergence

It is crucial to evaluate the convergence of DADPc, since it is an iterative algorithm. Figure 5 presents the average experimental results obtained from three task protocols of the SEED dataset, with the right y-axis indicating the values. The curves in the figure illustrate that the proposed algorithm exhibits asymptotic convergence. Overall, the objective values of DADPc stabilize within 60 iterations. This trend was also evident in other recognition tasks with varying cross-session configurations.

Figure 5 depicts the recognition error on the left y-axis. WUCE outperforms CUCE in recognition accuracy due to the latter's complexity. Throughout the iterations, we observed notable improvements in recognition accuracy. Although there are some fluctuations in WUCE's accuracy when the number of iterations exceeds 40, the recognition accuracy generally remains above 90%. After 130 iterations, CUCE's recognition accuracy exceeds 80%.

6.4.4 Effect of hyperparameters

As analyzed in Section 3.1 (Preliminary), the σ-norm approximates the $l_{2,1}$-norm as σ approaches 0 and converges to the Frobenius norm as σ tends to infinity. Figure 6 illustrates that the proposed method achieves optimal recognition performance at a σ value of 0.001 instead of these two extreme conditions. These results suggest that adaptively tuning σ during training could enhance the model's robustness by balancing sensitivity to outliers with feature representation fidelity.

We analyze the impact of hyper-parameters λ_1, λ_2, and λ_3 on the SEED datasets in Figure 7. The first subfigure in Figure 7 shows that performance varies with λ_1. λ_1 controls weight parameters. When λ_1 is above 0, performance improves, peaking at $\lambda_1 = 100$. The second subfigure in Figure 7 shows

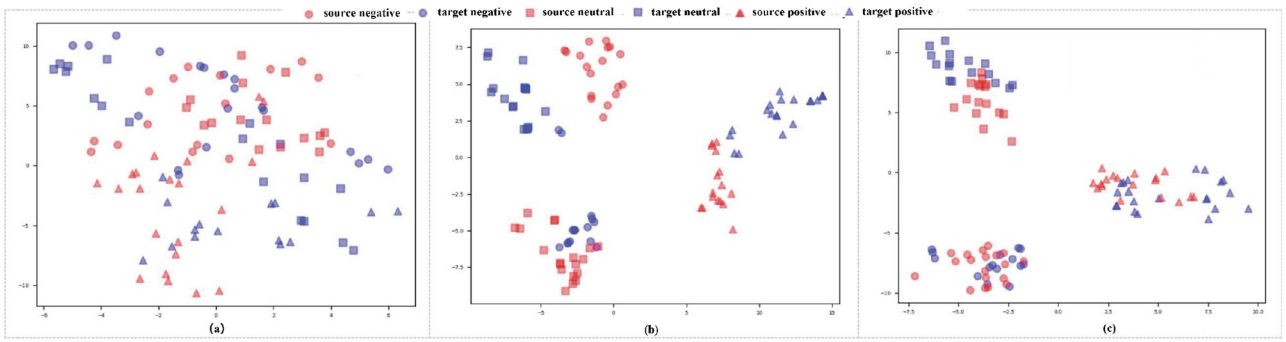

FIGURE 3
A visualization of the learned feature representations: **(a)** before training, **(b)** at the training epoch of 50, and **(c)** in the final model, respectively.

FIGURE 4
Confusion matrices of different models. **(a)** BiDANN Li et al. (2018b), **(b)** BiHDM Li et al. (2019c), **(c)** PR-PL Zhou et al. (2022), **(d)** DADPc.

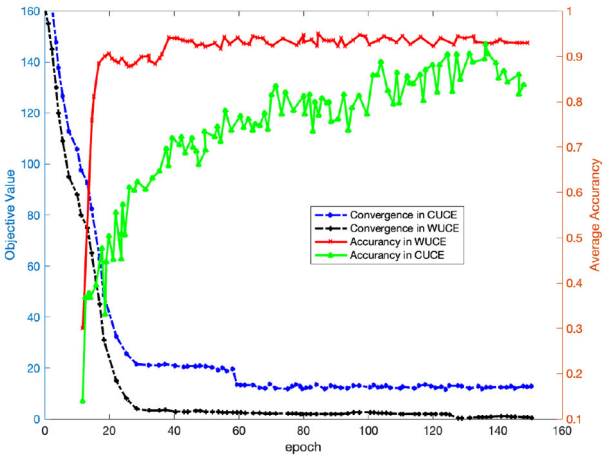

FIGURE 5
Convergence vs. accuracy on SEED.

the best performance at $\lambda_2 = 1$ or 10. As λ_1 and λ_2 near 0, performance drops, highlighting feature selection's role. The λ_1 and λ_2 prevent overfitting by regulating $\mathbf{W}$ and $\mathbf{b}$, crucial for the small-sized SEED datasets. Finally, we study λ_3, which regulates possibilistic clustering. The last subfigure in Figure 7 shows that when λ_3 is 10 or higher, performance is stable. Tuning λ_3 is difficult due to non-linearity (Nie et al., 2020). As λ_3 approaches 0, the DADPc's performance exhibits a significant decrease, showing that possibilistic clustering mitigates noisy data and improves generalization. These results show that each hyperparameter contributes to DADPC's adaptability and requires careful tuning for optimal performance in the CUCE scenario.

6.4.5 Ablation study

This ablation study systematically investigates the effectiveness of various components in the proposed model and presents the corresponding contributions of each component to the overall performance of the model DADPc. As shown in Table 8, when σ approaches 0, the adaptive norm of DADPc is the $\ell_{2,1}$ norm, and as σ approaches ∞, the adaptive norm of DADPc becomes the Frobenius (F) norm. It can be observed that as σ decreases, the performances of both CUCE and WUCE improve to varying degrees, while the performance of CUSE slightly declines. A possible reason is that the data differences generated by different experimental objects across different sessions for CUCE and WUCE may be greater than the data generated by the same experimental objects in various sessions for CUSE. The $\ell_{2,1}$ norm aids in feature selection, allowing it to identify more diverse and discriminative feature information from CUCE and WUCE, thus enhancing the model DADPc's discriminative effectiveness. However, a smaller σ does not always yield better results. A more detailed analysis of the hyper-parameter σ is provided in Section 6.4.5. Additionally, when $\lambda_1 = 0$, the constraint on W is removed, which can lead to model overfitting and hinder further feature selection, resulting in a decline in the model's performance. When $\lambda_2 = 0$, the constraint on b is removed, leading to only a slight decrease in the model's performance. Most importantly, when

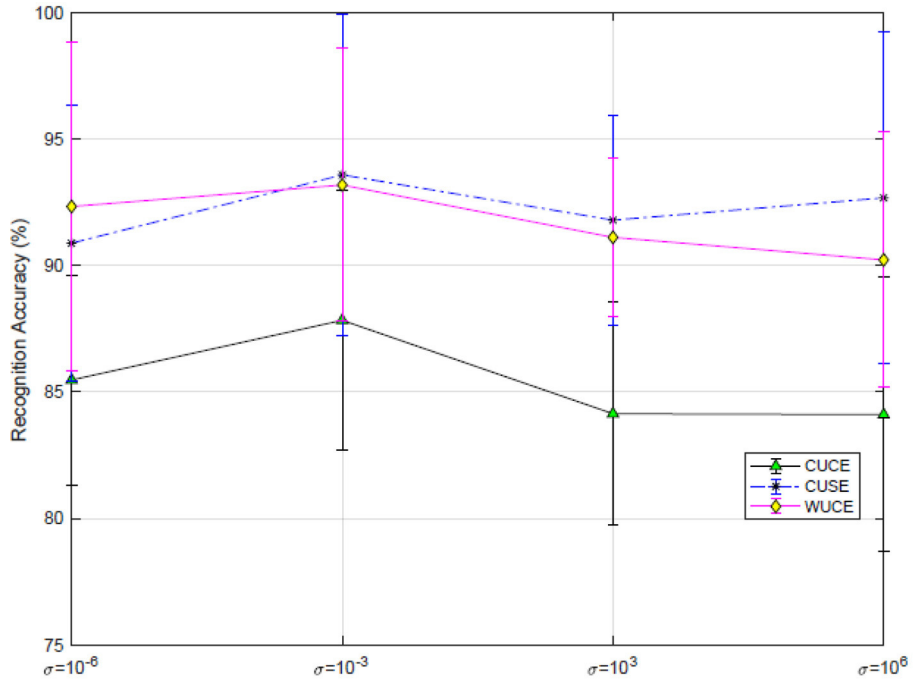

FIGURE 6
Effect of σ on SEED dataset.

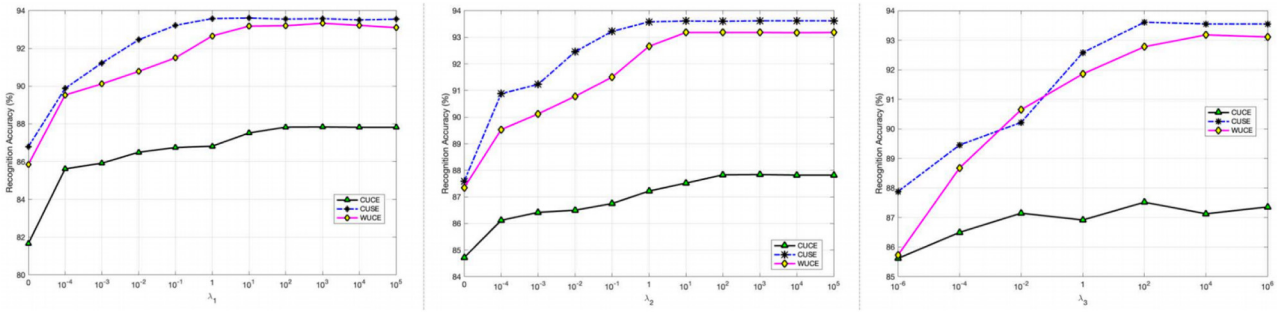

FIGURE 7
Effect of hyper-parameters on SEED dataset.

$\lambda_3 = 0$, it is equivalent to removing the possibilistic clustering constraint term. In these three different experimental scenarios, the model's performance drops by about 20% to 30%. This phenomenon indicates that the possibilistic clustering constraint term significantly impacts improving DADPc's performance.

6.4.6 Spatial mapping of key EEG patterns for emotion recognition

During the spatial mapping process, we pinpoint crucial neural activities associated with emotion recognition by assessing the mutual information shared between these patterns and the predictive labels. More precisely, during the ith validation phase, we work with a dataset from the target domain, denoted as X_i^t, which is structured as an $N_t * 310$ matrix. Here, N_t represents the number of data points in the target domain, while 310 corresponds to the DE features extracted from 62 electrodes across five frequency ranges. The model's output predictions for this phase are represented by $\hat{Y}_i^t$ with a size of $N_t * 3$. Each column in $\hat{Y}_i^t$ represents the three emotions (positive, neutral, and negative) along with their corresponding prediction probabilities (Ross, 2014). The resultant mutual information matrix, designated as $I(X_i^t, \hat{Y}_i^t) \in \mathbb{R}^{3\times 310}$, quantifies the intrinsic relationship between the EEG patterns and the model's predictions. This matrix $I\left(X_i^t, \hat{Y}_i^t\right)$ is then scaled to a range of [0,1], where higher values indicate a stronger contribution of the EEG patterns to the model's predictions during the ith validation phase. Figure 8 presents the average of all derived $I\left(X_i^t, \hat{Y}_i^t\right)$ matrices across various validation phases. As illustrated in Figure 8, the mutual information in the region indicated by the red arrow for the beta and gamma frequency bands exhibits

TABLE 8 The ablation accuracy(%) of our proposed model on SEED.

Ablation strategy	CUCE	CUSE	WUCE
DADPc with $\sigma = 10^{-8}$($l_{2,1}$-norm)	84.29 ± 4.37	90.24 ± 4.18	92.06 ± 6.25
DADPc with $\sigma = 10^{8}$(F-norm)	83.20 ± 5.26	90.37 ± 3.62	90.10 ± 4.39
DADPc w/o constraints on $\mathbf{W}(\lambda_1 = 0)$	85.49 ± 6.39	91.60 ± 5.72	91.06 ± 7.11
DADPc w/o constraints on $\mathbf{b}(\lambda_2 = 0)$	87.21 ± 3.28	93.08 ± 6.72	92.88 ± 6.55
DADPc w/o PC($\lambda_3 = 0$)	63.43 ± 7.14	62.89 ± 6.82	75.69 ± 7.18
DADPc	**87.83 ± 4.21**	**93.58 ± 6.35**	**93.18 ± 5.40**

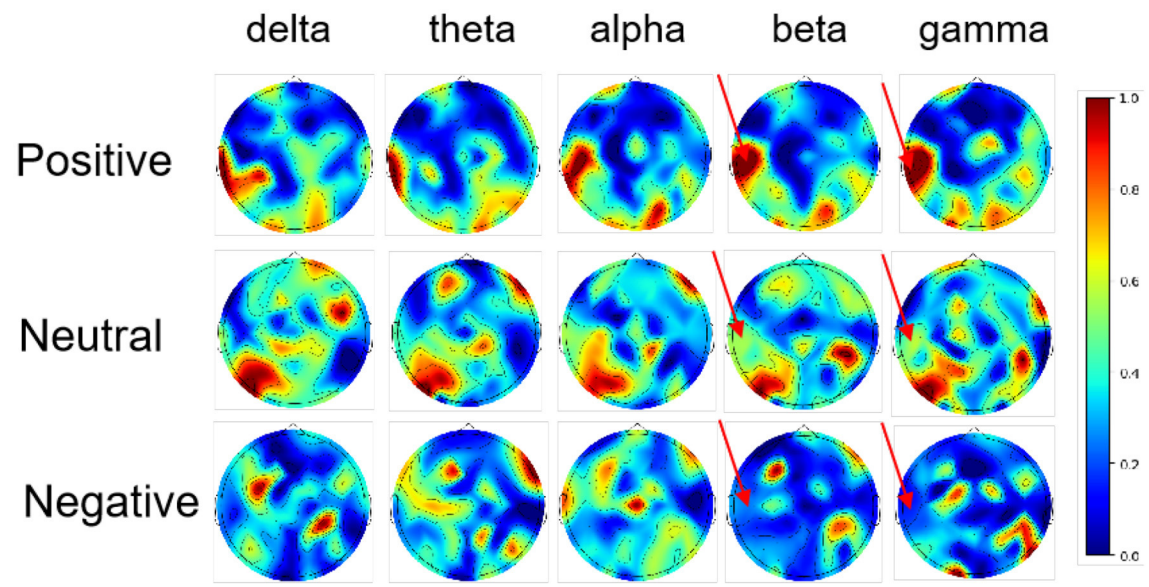

FIGURE 8
Topographic analysis of the mutual information between the EEG patterns and the model predictions.

a color gradient from red to blue, where the color corresponds to emotional intensity. This region is responsible for visual processing and emotional regulation, and its high-frequency neural activity is associated with the perception of complex emotional stimuli. A darker red tone signifies more intense emotions, while a bluer tone indicates calmer or less intense emotional states, which aligns with the findings of Zheng and Lu (2015), particularly in the region indicated by the red arrow.

7 Conclusion

This study presents Domain Adaptive Deep Possibilistic Clustering (DADPc), a novel framework that unifies deep domain-invariant feature learning and possibilistic clustering to address key challenges in EEG-based emotion recognition: inter-subject variability, label scarcity, and noise sensitivity. By reformulating maximum mean discrepancy (MMD) as a one-centroid clustering task within a fuzzy entropy-regularized possibilistic framework, DADPc mitigates noise-induced domain shifts while enhancing feature discriminability. The integration of adaptive weighted loss and memory bank strategies further enhances pseudo-label reliability and cross-domain alignment. Extensive experiments on SEED, SEED-IV, and DEAP datasets demonstrate DADPc's superiority. However, the manual tuning of λ_3 (fuzzy entropy weight) remains subjective, potentially limiting reproducibility across datasets (the last subfigure in Figure 7). Consequently, λ_3 and v_{ik} constitute a direction deserving further exploration through Bayesian optimization or a meta-learning strategy.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://epileptologie-bonn.de/cms/upload/workgroup/lehnertz/eegdata.html and http://bcmi.sjtu.edu.cn/seed.

Author contributions

YD: Writing – original draft. QL: Supervision, Writing – review & editing. XW: Data curation, Visualization, Writing – review & editing. DZ: Investigation, Validation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported in part by Zhejiang Basic Public Welfare Research Program under Grant ZCLTGY24F0301, in part by Opening Project of Ningbo Key Laboratory of Aging Health Equipment and Service Technology under Grant NZ25KF112, in part by Ningbo Natural Science Foundation Project under Grant 2022J180 and 2024J219, in part by open research fund of The Dazhou City Key Laboratory of Multidimensional Data Perception and Intelligent Information Processing under Grant DWSJ2404, and in part by Sichuan Provincial Engineering Research Center of Meteorological Photoelectric Sensing Technology and Applications under Grant 2024GCZX005, and in part by Special Polymer Materials for Automobile Key Laboratory of Sichuan Province under Grand TZGC2023ZB-05.

Acknowledgments

This is to acknowledge the Natural Science Foundation Committee of Zhejiang Province, the Ningbo Science and Technology Bureau, the Dazhou City Key Laboratory of Multidimensional Data Perception and Intelligent Information Processing, the Sichuan Provincial Engineering Research Center of Meteorological Photoelectric Sensing Technology and Applications, and the Institute of Artificial Intelligence Application at Ningbo Polytechnic for aiding the authors' efforts.

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.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

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

Alarco, S. M., and Fonseca, M. J. (2019). Emotions recognition using EEG signals: a survey. *IEEE Trans. Affect. Comp*. 10, 374–393. doi: 10.1109/TAFFC.2017.2714671

Baktashmotlagh, M., Harandi, M. T., Lovell, B. C., and Salzmann, M. (2013). "Unsupervised domain adaptation by domain invariant projection," in *2013 IEEE International Conference on Computer Vision* (Sydney, NSW: IEEE), 769–776.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Mach. Learn*. 79, 151–175. doi: 10.1007/s10994-009-5152-4

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

Bruzzone, L., and Marconcini, M. (2010). Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE Trans. Pattern Anal. Mach. Intell*. (5). doi: 10.1109/TPAMI.2009.57

Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E., and Bul, S. R. (2017). "Autodial: automatic domain alignment layers," in *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE). doi: 10.1109/ICCV.2017.542

Chen, H., Zhunan, L., Ming, J., and Jinpeng, L. (2021). "MEERNet: multi-source EEG-based emotion recognition network for generalization across subjects and sessions," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*(Mexico: IEEE), 6094–6097.

Chen, Z., Zhuang, J., Liang, X., and Lin, L. (2019). "Blending-target domain adaptation by adversarial meta-adaptation networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA), 2243–2252. doi: 10.1109/CVPR.2019.00235

Chu W.-S., De la Torre, F., and Cohn, J. F. (2013). "Selective transfer machine for personalized facial action unit detection," in *Proceedings of 2013 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Portland, OR: IEEE), 3515–3522.

Dan, Y., Di, Z., and Wang, Z. (2024). Discriminative possibilistic clustering promoting cross-domain emotion recognition. *Front. Neurosci*. 18:1458815. doi: 10.3389/fnins.2024.1458815

Dan, Y., Tao, J., Fu, J., and Zhou, D. (2021). Possibilistic clustering-promoting semi-supervised learning for EEG-based emotion recognition. *Front. Neurosci*. 15:690044. doi: 10.3389/fnins.2021.690044

Dan, Y., Tao, J., and Zhou, D. (2022). Multi-model adaptation learning with possibilistic clustering assumption for eeg-based emotion recognition. *Front. Neurosci*. 16:855421. doi: 10.3389/fnins.2022.855421

Ding, Z., Li, S., Shao, M., and Fu, Y. (2018). "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *European Conference on Computer Vision 2018* (Munich), 36–52. doi: 10.1007/978-3-030-01216-8_3

Dino, S., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist*. 41, 2263–2291. doi: 10.1214/13-AOS1140

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," in *Proceedings of the 31st International Conference on International Conference on Machine Learning* (New York: JMLR), I-647–655.

Du, X., Lai, Y. K., Deng, X., Wang, H., Ma, C., Li, J., et al. (2022). An efficient lstm network for emotion recognition from multichannel EEG signals. *IEEE Trans. Affect. Comp*. 13, 1528–1540. doi: 10.1109/TAFFC.2020.3013711

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. New York: Wiley-Interscience.

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. (2013). "Unsupervised visual domain adaptation using subspace alignment," in *2013 IEEE International Conference on Computer Vision* (Sydney, NSW: IEEE), 2960–2967.

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. doi: 10.1007/978-3-319-58347-1_10

Ghifary, M., Balduzzi, D., Klejin, B., and Zhang, M. (2017). Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Trans. Pattern Anal. Mach. Intell*. 99, 1–1. doi: 10.1109/TPAMI.2016.2599532

Gong, B., Yuan, S., Fei, S., and Kristen, G. (2012). "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI: IEEE), 2066–2073.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). "A fast, consistent kernel two-sample test," in *Advances in Neural Information Processing*

Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a Meeting held 7–10 December 2009 (Vancouver, BC: DBLP).

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

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

Jin, J., Wang, Z., Xu, R., Liu, C., Wang, X., and Cichocki, A. (2023). Robust similarity measurement based on a novel time filter for ssveps detection. *IEEE Trans. Neural Netw. Learn. Syst*. 34, 4096–4105. doi: 10.1109/TNNLS.2021.3118468

Jin, Y.-M., Luo, Y.-D., Zheng, W.-L., and Lu, B.-L. (2017). "EEG-based emotion recognition using domain adaptation network," in *2017 International Conference on Orange Technologies (ICOT)* (Singapore: IEEE), 222–225.

Kang, G., Jiang, L., Wei, Y., Yang, Y., and Hauptmann, A. (2022). 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

Kim, M.-K., Kim, M., Oh, E., and Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from 13 the human EEG. *Comput. Math. Methods Med*. 2013:573734. doi: 10.1155/2013/573734

Koelstra, S., Patras, I., Muhl, C., Nijholt, A., Soleymani, M., Pun, T., et al. (2012). DEAP: A database for emotion analysis; using physiological signals. *IEEE Trans. Affective Comp*. 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Krishnapuram, R., and Keller, J.-M. (1993). A possibilistic approach to clustering. *IEEE Trans. Fuzzy Syst*. 1, 98–110. doi: 10.1109/91.227387

Lan, Z., Sourina, O., Wang, L., Scherer, R., and Muller-Putz, G. R. (2019). Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Trans. Cognit. Dev. Syst*. 11, 85–94. doi: 10.1109/TCDS.2018.2826840

Laurens, V. D. M., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res*. 9, 2579–2605. doi: 10.48550/arXiv.2108.01301

Lee, D. H. (2013). Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. *Comput. Sci.* Available online at: https://api.semanticscholar.org/CorpusID:18507866

Lee, S., Kim, D., Kim, N., and Jeong, S.-G. (2019). "Drop to adapt: learning discriminative features for unsupervised domain adaptation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 91–100.

Li, J., Qiu, S., Shen, Y.-Y., Liu, C.-L., and He, H. (2019a). Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Trans. Cybern*. 50, 3281–3293. doi: 10.1109/TCYB.2019.2904052

Li, J., Shuang, Q., Changde, D., Yixin, W., and Huiguang, H. (2019b). Domain adaptation for EEG emotion recognition based on latent representation similarity. *IEEE Trans. Cognit. Dev. Syst*. 12, 344–353. doi: 10.1109/TCDS.2019.2949306

Li, X., Hu, B., Sun, S., and Cai, H. (2016). EEG-based mild depressive detection using feature selection methods and classifiers. *Comput. Methods Programs Biomed*. 136, 151–161. doi: 10.1016/j.cmpb.2016.08.010

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

Li, Y., Zhang, W., Zong, Y., Cui, Z., Zhang, T., and Zhou, X. (2018b). A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comp*. 12, 494–504. doi: 10.1109/TAFFC.2018.2885474

Li, Y., Zheng, W., Cui, Z., Zong, Y., and Ge, S. (2018). EEG emotion recognition based on graph regularized sparse linear regression. *Neural Proc. Lett*. 49, 555–571. doi: 10.1007/s11063-018-9829-1

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. Comp*. 13, 568–578. doi: 10.1109/TAFFC.2019.2922912

Li, Y., Zheng, W., Wang, L., Zong, Y., and Cui, Z. (2022). From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comp*. 13, 568–578.

Li, Y., Zheng, W., Wang, L., Zong, Y., Qi, L., Cui, Z., et al. (2019c). *A Novel Bi-Hemispheric Discrepancy Model for EEG Emotion Recognition*.

Liang, J., Ran, H., Zhenan, S., and Tieniu, T. (2019). Aggregating randomized clustering-promoting invariant projections for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*. 41, 1027–1042. doi: 10.1109/TPAMI.2018.2832198

Litrico, M., Del Bue, A., and Morerio, P. (2023). "Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC: IEEE), 7640–7650.

Long, M., Cao, Y., Cao, Z., Wang, J., and Jordan, M. I. (2019). Transferable representation learning with deep adaptation networks. *IEEE Trans. Pattern Anal. Mach. Intell*. 41, 3071–3085. doi: 10.1109/TPAMI.2018.2868685

Long, M., Jianmin, W., Guiguang, D., Jiaguang, S., and Yu, P. S. (2013). "Transfer feature learning with joint distribution adaptation," in *2013 IEEE International Conference on Computer Vision* (Sydney, NSW: IEEE), 2200–2207. doi: 10.1109/ICCV.2013.274

Long, M., Wang, J., Sun, J., and Yu, P. S. (2015). Domain invariant transfer kernel learning. *IEEE Trans. Knowl. Data Eng*. 27, 1519–1532. doi: 10.1109/TKDE.2014.2373376

Long, M., Zhu, H., Wang, J., and Jordan, M. (2016). "Unsupervised domain adaptation with residual transfer networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16* (Red Hook, NY: Curran Associates Inc), 136–144.

Luo, T., Zhang, J., Qiu, Y., Zhang, L., Hu, Y., Yu, Z., et al. (2024). *MDDD: Manifold-Based Domain Adaptation with Dynamic Distribution for Non-Deep Transfer Learning in Cross-Subject and Cross-Session EEG-Based Emotion Recognition*.

Luo, Y., Zhang, S.-Y., Zheng, W.-L., and Lu, B.-L. (2018). "Wgan domain adaptation for EEG-based emotion recognition," in *Neural Information Processing*, eds. L. Cheng, A. C. S. Leung, and S. Ozawa (Cham: Springer International Publishing), 275–286.

Ma, B.-Q., Li, H., Zheng, W.-L., and Lu, B.-L. (2019). "Reducing the subject variability of EEG signals with adversarial domain generalization," in *Neural Information Processing*, eds. T. Gedeon, K. W., Wong, and M. Lee (Cham: Springer International Publishing), 30–42.

Magdiel, J.-G., and Gibran, F.-P. (2023). Cross-subject eeg-based emotion recognition via semisupervised multisource joint distribution adaptation. *IEEE Trans. Instrum. Meas*. 72, 1–11. doi: 10.1109/TIM.2023.3302938

Muhl, C., Allison, B., Nijholt, A. and Chanel, G. (2014). A survey of affective brain computer interfaces:principles, state-of-the-art, and challenges. *Brain-Comp. 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., Dong, X., Tian, L., Wang, R., and Li, X. (2020). Unsupervised feature selection with constrained 2,0-norm and optimized graph. *IEEE Trans. Neural Netw. Learn. Syst*. 33, 1702–1713. doi: 10.1109/TNNLS.2020.3043362

Nie, F., Wang, H., Huang, H., and Ding, C. (2013). "Adaptive loss minimization for semi-supervised elastic embedding," in *2013 Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI-2013)* (Beijing: Department of Computer Science and Engineering University of Texas at Arlington and Department of Electrical Engineering and Computer Science Colorado School of Mines), 1565–1571.

Pan S. J., and Yang, Q. (2018). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*., 22:1345–1359. doi: 10.1109/TKDE.2009.191

Pan, S. J., Tsang, I. W., Kwok, J., 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. R. (2019). "Emotional state recognition with eeg signals using subject independent approach," in *Data Science and Big Data Analytics* (Cham: Springer), 117–124.

Patel, 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

Peng, Y., Liu, H., Li, J., Huang, J., Lu, B.-L., and Kong, W. (2023). Cross-session emotion recognition by joint label-common and label-specific eeg features exploration. *IEEE Trans. Neural Syst. Rehabil. Eng*. 31, 759–768. doi: 10.1109/TNSRE.2022.3233109

Peng, Y., Wang, W., Kong, W., Nie, F., Lu, B.-L., and Cichocki, A. (2022). Joint feature adaptation and graph adaptive label propagation for cross-subject emotion recognition from EEG signals. *IEEE Trans, Affect. Comp*. 13, 1941–1958. doi: 10.1109/TAFFC.2022.3189222

Philip, H., Frerix, T., Mordvintsev, A., and Cremers, D. (2017). "Associative domain adaptation," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE), 2784–2792. doi: 10.1109/ICCV.2017.301

Pinheiro, P. O. (2018). "Unsupervised domain adaptation with similarity learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 8004–8013.

Ronan, C., Sinz, F., Weston, J., and Bottou, L. (2006). Large scale transductive svms. *J. Mach. Learn. Res*. 7, 1687–1712. doi: 10.1007/s10846-006-9063-3

Ross, B. C. (2014). Mutual information between discrete and continuous datasets. *PLoS ONE* 9:e87357. doi: 10.1371/journal.pone.0087357

Saito, K., Kim, D., Sclaroff, S., and Saenko, K. (2020). Universal domain adaptation through self supervision. *arXiv [Preprint]*. doi: 10.48550/arXiv.2002.07953

Schlkopf, B., Smola, A., and Muller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput*. 10, 1299–1319. doi: 10.1162/089976698300017467

Shi, L.-C., Jiao, Y.-Y., and Lu, B.-L. (2013). "Differential entropy feature for EEG-based vigilance estimation," in *2013 35th Annual International Conference*

of the IEEE Engineering in Medicine and Biology Society (EMBC) (Osaka: IEEE), 6627–6630.

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comp*. 11, 532–541. doi: 10.1109/TAFFC.2018.2817622

Stern, P. R. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194. doi: 10.1126/science.1076358

Sun, B., Feng, J., and Saenko, K. (2015). "Return of frustratingly easy domain adaptation," in *AAAI Conference on Artificial Intelligence. CoRR, 2015*. doi: 10.48550/arXiv.1511.05547

Sun, B., and Saenko, K. (2016). *Deep Coral: Correlation Alignment for Deep Domain Adaptation*.

Tang, H., and Jia, K. (2019). "Discriminative adversarial domain adaptation," in *Association for the Advancement of Artificial Intelligence (AAAI) 2020*, 5454–6242. doi: 10.1609/aaai.v34i04.6054

Tao, J., Chung, F. L., and Wang, S. (2012). On minimum distribution discrepancy support vector machine for domain adaptation. *Pattern Recognit*. 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., 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

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

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

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

Tao, J., Yan, L., and He, T. (2024). Domain-invariant adaptive graph regularized label propagation for eeg-based emotion recognition. *IEEE Access* 12:3454082. doi: 10.1109/ACCESS.2024.3454082

Tao, J., Zhou, D., Liu, F., and Zhu, B. (2019). Latent multi-feature co-regression for visual recognition by discriminatively leveraging multi-source models. *Pattern Recognit*. 87:296–316. doi: 10.1016/j.patcog.2018.10.023

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. arXiv [preprint] arXiv:1412.3474. doi: 10.48550/arXiv.1412.3474

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Vinay, J., Alamgir, M., Altun, Y., Schlkopf, B., and Grosse-Wentrup, M. (2015). Transfer learning in brain-computer interfaces. *IEEE Comp. Intellig. Magazine* 11, 20–31. doi: 10.1109/MCI.2015.2501545

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., et al. (2023). Generalizing to unseen domains: a survey on domain generalization. *IEEE Trans. Knowl. Data Eng*. 35, 8052–8072. doi: 10.48550/arXiv.2103.03097

Wei, W., Zhang, Y., Zhu, J., and Lu, B. (2015). "Transfer components between subjects for eeg-based emotion recognition," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xi'an,: IEEE), 917–922. doi: 10.1109/ACII.2015.7344684

Wenming, Z. (2017). Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis. *IEEE Trans. Cognit. Dev. Syst*. 9, 281–290. doi: 10.1109/TCDS.2016.2587290

Zhang, G., and Etemad, A. (2022). PARSE: pairwise alignment of representations in semi-supervised EEG learning for emotion recognition. *arXiv* [preprint] arXiv:2202.05400. doi: 10.1109/TAFFC.2022.3210441

Zhang, Y., Dong, J., Zhu, J., and Wu, C. (2019). 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., Tian, F., Wu, H., Geng, X., Qian, D., Dong, J., et al. (2013). Brain mri tissue classification based fuzzy clustering with competitive learning. *Med. Imag. Health Inform*. 7, 1654–1659. doi: 10.1166/jmihi.2017.2181

Zheng, W.-L., and Lu, B.-L. (2013). "Personalizing EEG-based affective models with transfer learning," in *2016 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI–16)* (New York, NY), 2732–2738.

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

Zhong, P., Wang, D., and Miao, C. (2020). *EEG-Based Emotion Recognition Using Regularized Graph Neural Networks*.

Zhong, P., Wang, D., and Miao, C. (2022). EEG-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comp*. 13, 1290–1301. doi: 10.1109/TAFFC.2020.2994159

Zhou, R., Ye, W., Zhang, Z., Luo, Y., Zhang, L., Li, L., et al. (2023). EEGmatch: learning with incomplete labels for semi-supervised EEG-based cross-subject emotion recognition. *IEEE Trans. Neural Netw. Learn. Syst*. 36, 12991–13005. doi: 10.1109/TNNLS.2024.3493425

Zhou, R., Zhang, Z., Yang, X., Fu, H., Zhang, L., Li, L., et al. (2022). Pr-pl: A novel prototypical representation based pairwise learning framework for emotion recognition using EEG signals. *IEEE Trans. Affect. Comp*. 72, 1–11. doi: 10.1109/TAFFC.2023.3288118

Zhu, J., Zou, H., Rossett, S., and Haster, T (2006). Multi-class adaboost. *Stat. Interface* 2, 349–360. doi: 10.4310/SII.2009.v2.n3.a8

Zhu, Y., Zhuang, F., Wang, J., Ke, G., Chen, J., Bian, J., et al. (2021). Deep subdomain adaptation network for image classification. *IEEE Trans. Neural Netw. Learn. Syst*. 32, 1713–1722. doi: 10.1109/TNNLS.2020.2988928