Multi-Modal Cross-Subject Emotion Feature Alignment and Recognition With EEG and Eye Movements

Qi Zhu, *Member, IEEE*, Ting Zhu, Lunke Fei, *Senior Member, IEEE*, Chuhang Zheng, Wei Shao, David Zhang, *Life Fellow, IEEE*, and Daoqiang Zhang, *Senior Member, IEEE*

***Abstract*—Multi-modal emotion recognition has attracted much attention in human-computer interaction, because it provides complementary information for the recognition model. However, the distribution drift among subjects and the heterogeneity of different modalities pose challenges to multi-modal emotion recognition, thereby limiting its practical application. Most of the current multi-modal emotion recognition methods are difficult to suppress above uncertainties in fusion. In this paper, we propose a cross-subject multi-modal emotion recognition framework, which jointly learns subject-independent representation and common feature between EEG and eye movements. First, we design the dynamic adversarial domain adaptation for cross-subject distribution alignment, dynamically selecting source domains in training. Second, we simultaneously capture intra-modal and inter-modal emotion-related features by both self-attention and cross-attention mechanisms, thus obtaining the robust and complementary representation of emotional information. Then, two contrastive loss functions are imposed on above network to further reduce inter-modal heterogeneity, and mine higher-order semantic similarity between synchronously collected multi-modal data. Finally, we used the output of the softmax layer as the predicted value. The experimental results on several multi-modal emotion datasets with EEG and eye movements demonstrate that our method is significantly superior to the state-of-the-art emotion recognition approaches.**

***Index Terms*—EEG, cross-subject, multi-modal fusion, emotion recognition, contrastive learning, adversarial domain adaptation.**

Received 8 August 2024; revised 19 March 2025; accepted 19 March 2025. Date of publication 24 March 2025; date of current version 15 September 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62371234, Grant 62076129, Grant 62136004, Grant 62272226 and Grant 62176066, and in part by the Key Research and Development Plan of Jiangsu Province under Grant BE2022842. Recommended for acceptance by D. Zhang. *(Qi Zhu and Ting Zhu equally contributed to this work.) (Corresponding authors: Wei Shao; Daoqiang Zhang.)*

Qi Zhu, Ting Zhu, Chuhang Zheng, Wei Shao, and Daoqiang Zhang are with the College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China, and also with Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: zhuqinuaa@163.com; zhuting@nuaa.edu.cn; h.zheng@nuaa.edu.cn; shaowei20022005@nuaa.edu.cn; dqzhang@nuaa.edu.cn).

Lunke Fei is with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: flksxm@126.com).

David Zhang is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: davidzhang@cuhk.edu.cn).

Our code is available at: https://github.com/xbrainnet/CSMM.

Digital Object Identifier 10.1109/TAFFC.2025.3554399

I. Introduction

EMOTION recognition plays an important role in understanding human intention and behavior, and it is one of the core research topics in human-computer interaction [1]. Since emotions are affected by multiple complex factors, such as physiology, psychology, and environment, it is often difficult for machines to accurately understand human emotions. In recent years, many intelligent emotion recognition methods have been proposed [2], [3]. However, the uncertainty of emotion data, mainly reflected in modal heterogeneity and cross-subject aspects, restricts the application of emotion recognition in practice.

There are a variety of emotion recognition methods based on behavioral or physiological traits. Among the physiological traits, Electroencephalogram (EEG) is the most widely used modality in emotion recognition due to its high temporal resolution and anti-cheating. For EEG feature extraction, LSTM and CNN have been used to directly extract global features and local features from EEG signal [4]. Xing et al. [5] used a stacked autoencoder (SAE) to model the correlation between EEG channels and considerably decompose the EEG source signals from the collected EEG signals. Long short-term memory (LSTM) networks are employed to extract temporal emotion-related information from differential entropy (DE) features [6]. However, the EEG typically exhibits a low signal-to-noise ratio, which constrains the accuracy of emotion recognition when relying on this single modality.

As multi-modal fusion can provide complementary semantic information for emotion recognition, many studies have combined EEG with behavioral data, such as eye movements and facial expressions, to improve the performance of emotion recognition. For example, Huang et al. [7] projected EEG and facial expression modalities into common feature space with canonical correlation analysis. It is worth noting that facial expressions can be more easily influenced by external factors, such as lighting, context, and culture. Specifically, the meaning of a facial expression can vary significantly across different cultures. Compared to facial expressions, eye movement is more consistent across individuals and cultures, and it has received more and more attention in the field of emotion recognition. Lu et al. [8] performed feature level fusion by concatenating feature matrices of eye movement and EEG modalities, and adopted a support vector machine with the linear kernel for

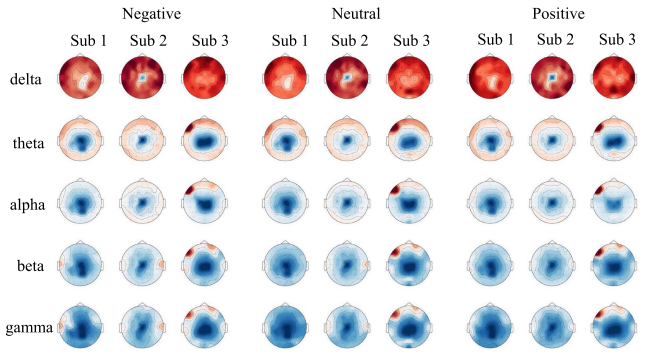


Fig. 1. The neural patterns of DE features in five frequency bands from different subjects. Areas with red color indicate the high energy, and areas with black color indicate the low energy.

classification. Liu et al. [9] proposed a deep CCA-based method with an attention mechanism to improve the accuracy of emotion recognition based on EEG and eye movement modalities. Liu et al. [10] performed the experiments of bi-modal deep auto-encoder (BDAE) [11] with EEG and eye movement modalities, in which two restricted Boltzmann machines (RBMs) were constructed for these two modalities, and another RBM is stacked to perform the further fusion. However, the problem of effectively integrating heterogeneous EEG and eye movement modalities still remains an open challenge in emotion recognition.

In real-world applications, distributional differences between individuals pose challenges to multi-modal emotion recognition. As shown in Fig. 1, different subjects had obvious individual inherent characteristics, which would interfere with the accuracy of emotion recognition. Specifically, the phenomenon that differences in data distribution between subjects are huger than that between emotion categories may lead to misclassification, and the heterogeneity of multi-modal data further aggravates the complexity of the above problem. Recently, the cross-subject issue has been paid more and more attention in emotion recognition, and some methods for cross-subject emotion recognition have been proposed [12]. However, these existing methods are specifically designed for subject alignment with a single modality and cannot effectively alleviate the cross-subject problem in multi-modal emotion recognition. Therefore, there is an urgent need to investigate and develop effective multimodal emotion recognition methods that can accurately identify emotions across different subjects.

In this study, we propose a robust cross-subject multi-modal emotion recognition method. First, a self-paced adversarial domain adaptation network is developed to capture the subject-independent emotion-related embeddings against distributional differences among subjects. It simulates human cognition mechanisms by dynamically adding subjects to the training process, which can avoid the negative effects of hard learning subjects on the initial model. Second, we adopt both self-attention and cross-attention mechanisms to extract the intra-modal and inter-modal feature representations related to emotion. Moreover, by mining semantic similarity in multi-modal data, we designed two contrastive loss functions from the feature level and sample level, respectively, to promote modal alignment and fusion. Specifically, at the feature level, we construct the contrastive loss between features from different modalities to improve the cross-modal alignment before fusion. At the sample level, we construct the contrastive loss between sample pairs composed of synchronous and asynchronous modalities to capture inter-modal complementary information. Finally, the classification results can be obtained by feeding the multi-modal fused representation into a linear layer and a softmax layer. The proposed method is significantly different from the existing method, such as [13]. First, there are differences in the learning tasks. The existing method [13] is based on EEG data for cross-subject emotion recognition, while our method accomplishes cross-subject emotion recognition tasks based on a variety of heterogeneous modalities including EEG and eye movements. Therefore, the problem we are studying is more challenging. In actual applications, multi-modal methods can extract complementary emotional information, which usually leads to better recognition effects than unimodal methods. Second, the objectives of contrastive learning in these two works differ significantly. The existing method [13] uses contrastive learning to extract subject-invariant features from EEG data, whereas our method employs adversarial domain adaptation to extract domain-invariant representations and adopts contrastive learning to extract inter-modality fusion features on the basis of EEG and eye movement modalities. Third, the construction methods of the contrast tasks are different. The existing method [13] forms positive sample pairs from EEG samples of the same trial across different subjects and negative sample pairs from EEG samples of different trials. In contrast, our method constructs contrastive tasks both at the feature level and the sample level. At the feature level, we maximize inter-modality similarity by constructing a modal contrast loss between EEG and eye movement modalities, paving the way for subsequent fusion of modal features. At the sample level, we form positive sample pairs from synchronously collected EEG and eye movement modal samples and negative sample pairs from asynchronously collected EEG and eye movement modal samples to learn high-order emotional information.

In brief, our primary contributions are summarized as follows:

1) We propose a multi-modal emotion recognition method with both cross-subject alignment and feature fusion between EEG and eye movements. We investigate both above challenges of emotion recognition in a unified framework, where the solutions to these challenges are mutually reinforcing, thereby improving the performance of emotion recognition.
2) We propose a dynamic cross-subject alignment method, in which a self-paced learning strategy is introduced to the adversarial domain adaptation network. It dynamically selects the easy-learning subjects and gradually adds them to the training process to alleviate the negative impact of domain distributional variation across subjects.
3) The self-attention mechanism and cross-attention mechanism are adopted to extract the intra-modal and inter-modal emotion-related features. Simultaneously, we improve multi-modal alignment and fusion by constructing two contrastive losses at feature level and sample level,

respectively, and capture the higher-order complementary features between modalities.

The remainder of this paper is organized as follows: In Section II, we elaborated on the related works. Then our proposed cross-subject multi-modal emotion recognition method is presented in Section III. Section IV describes the materials and the details of the experiment. In Section V, we provide the experimental results and discussion. The paper is concluded in Section VI.

II. Related Work

A. Multi-Source Domain Adaptation

Subject-independent representations can be obtained by domain adaptation to reduce the limitation of huge original distribution differences among individuals. The domain adaptation techniques attempt to project the data from the original space into a common feature space, in which the distribution difference is eliminated as far as possible through a nonlinear mapping function. For example, transfer component analysis [14] maps data from two domains into a reproducing kernel hilbert space (RKHS), in which the distribution distance between source and target domains is reduced. Transductive parameter transfer method [15] is designed for multi-source domain adaptation, and it predicts the parameters of the target classifier according to the learned relationship between the data distributions and the parameter vectors of the classifier constructed for each source domain. In addition, deep neural networks are also used to learn nonlinear mappings. Gratton et al. [16] proposed to embed maximum mean discrepancy (MMD) in neural networks as the measure of distance between two domains, so that the distribution of source and target domain data is similar in the high-dimensional common projection space.

In recent years, adversarial approaches have shown remarkable performance in domain adaptation. Domain adversarial neural network (DANN) [17] adopts domain classifier and gradient reversal layer to guide encoder layers to generate domain-invariant representation. In DANN, feature extractor, label predictor, and domain classifier are denoted as G_f, G_y, and G_d, respectively. We define the input as $X \in \mathbb{R}^{n\times m}$, where n denotes the number of samples and m denotes the dimension. G_f and G_y are trained to minimize the classification loss on source domain:

$$L_y = H(G_y(G_f(X)), y) = -\sum_{i=1}^{M} y_i log(G_y(G_f(X))) \quad (1)$$

where M is the number of categories to recognize, and y is the real emotion label. Domain classifier G_d is a binary classifier that predicts whether a sample comes from the source domain or the target domain. The loss of domain classifier is defined by

$$L_d = H(G_d(G_f(X)), d) \quad (2)$$

where $d \in \{0, 1\}$ is the real domain label. Inspired by the generative adversarial networks (GANs), the parameters of G_d are updated by minimizing L_d, while the parameters of G_f are updated by maximizing L_d.

In addition, Pei et al. [18] refined the domain alignment for class alignment and proposed multi-adversarial domain adaptation (MADA). Tzeng et al. [19] introduced adversarial discriminative domain adaptation (ADDA), which learns a discriminative representation using the labeled source domain data and then projects the target data into a common space using an asymmetric mapping learned through domain adversarial loss. It is worth noting that DANN was originally proposed for the single-source domain problem, and then applied to the multi-source domain adaptation problem by treating multiple source domains as one entire source domain simply. Zhao et al. [20] proposed to train an independent feature extractor for each source domain, and then utilized Wasserstein distance to select samples close to the target domain to finetune the classifier. Zhao et al. [21] proposed multi-source domain adversarial networks (MDAN), which optimized task-adaptive generalization bounds to deal with multi-source domain adaptation. Peng et al. [22] integrated domain invariant feature learning, emotional state estimation, and optimal graph learning into a unified framework to propose a joint feature adaptive and graph adaptive label propagation model (JAGP) for cross-subject emotion recognition based on EEG signals. Magdiel et al. [23] presented a semi-supervised domain adaptation framework (FSA-TSP) to align the joint distributions of subjects. She et al. [24] considered both domain invariant and domain-specific features, and proposed multi-source transfer network (MS-ADA) based on multi-source associate domain adaptation network.

B. Correlation Mining in Multi-Modal Fusion

Multi-modal fusion aims to combine the data of multiple modalities and extract complementary information to facilitate downstream prediction tasks. The subspace learning-based fusion methods, such as CCA-based methods and MKL, try to project different modalities into a common subspace by maximizing the correlation across modalities. CCA [25] and its variants, such as Kernel CCA (KCCA) [26], DCCA [27] perform multi-modal alignment before fusion by maximizing the correlation among different modalities, while the modal-specific kernel in MKL [28] can be combined for heterogeneous data.

The attention mechanism has also been widely used in the field of multi-modal fusion recently. Wei et al. [29] stacked the features of different modalities as a whole input in cross attention module. Pan et al. [30] proposed a multi-modal attention mechanism to model the correlation among three modalities. Arsha et al. [31] introduced a transformer-based network architecture, which uses attention bottlenecks for modality fusion at multiple layers.

Contrastive learning has been gradually applied to multi-modal fusion in recent years. Jia et al. [32] used a dual-encoder architecture with a contrastive loss to align representations of image and text modalities. Li et al. [33] introduced the contrastive loss to align the image and text representation before multi-modal fusion to achieve better unimodal representation, and induced the modality matching loss to improve the robust predictions. Additionally, contrastive learning is also applied to cross-subject emotion recognition based on EEG signals due to

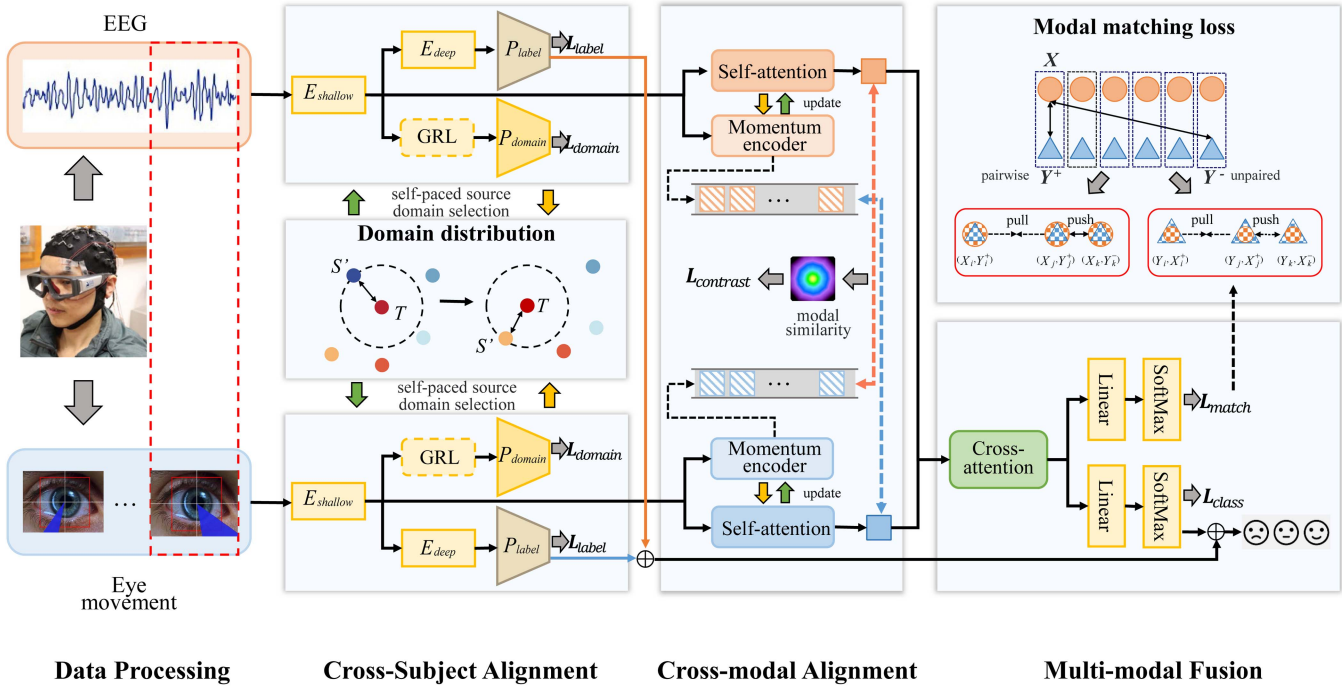


Fig. 2. The model architecture of the proposed cross-subject multi-modal emotion recognition method.

its ability to identify and utilize similarities across different data representations. For example, Li et al. [34] integrated convolutional neural network with supervised contrastive learning to enhance the classification accuracy and discrimination of emotional states by reducing model parameters and computational burden. Shen et al. [13] leveraged contrastive learning to minimize inter-subject variability in EEG data, thereby improving cross-subject emotion recognition performance.

Recently, some methods have been proposed for learning subject-invariant emotional feature representations from multimodal data. Magdiel et al. [35] proposed the CFDA-CSF method for cross-subject emotion recognition utilizing EEG signals and eye movement data, which enhances knowledge transfer between subjects by aligning coarse and fine-grained feature distributions and promoting inter-class separability in the target domain. Chen et al. [36] proposed an integrated multi-source learning network (CMSLNet) that balances the consistency and diversity of heterogeneous modalities in cross-subject emotion recognition. Gong et al. [37] proposed a multimodal emotion recognition model (CiABL) that enhances generalization across subjects by capturing complete modality representations from EEG and eye movement signals, aligning both marginal and conditional distributions to reduce individual variability, and offering model interpretability to understand the contribution of different features in recognizing emotions.

III. Proposed Method

In this section, we first introduce the model architecture of the proposed cross-subject multi-modal emotion recognition framework (CSMM) for EEG and eye movement signals and then specify its modules in detail. We draw an overview of the model architecture of CSMM in Fig. 2. The network is divided into three parts, which are cross-subject alignment, cross-modal alignment, and multi-modal fusion. The DE feature vectors from multiple frequency bands of EEG modality and statistical feature vectors of eye movement modality are extracted as the input of the framework in the stage of data processing.

A. Cross-Subject Alignment

In order to reduce the negative effects of the distribution drift among subjects, we develop a self-paced adversarial domain adaptation network to extract domain-invariant features. Specifically, the data from the same subject are treated as a domain, and the data are divided into the source and target domain data according to the visibility of the labels in the training process. Most of the previous adversarial domain adaptation networks are specifically designed for single-source domain adaptation problems, and such methods treat multiple source domains as one whole source domain in dealing with multi-source domain adaptation problems. This ill-considered approach ignores the semantic difference among source domains and affects the stability of the model during training. To alleviate the above problems, we introduced self-paced learning, which is a human being strategy, to gradually convert the multi-source domain adaptation problem into the single-source domain adaptation problem, alleviating the negative effects of hard-learning subjects to the initial model.

The adversarial domain adaptation network aims to motivate the feature encoder to generate domain-invariant and class-discriminative representations in the feature space. The training goals of the domain predictor and the feature encoder are competitive, the former guides the separation of source

and target domains, while the latter aims at the confusion. In addition, the emotion label predictor is added after the feature encoder to ensure the class discrimination of generated representation. We define the input as $X \in \mathbb{R}^{n\times m}$, where n denotes the number of samples and m denotes the dimension. Considering that the shallow layers of the feature encoder are prone to generate domain-invariant features while deep layers tend to generate downstream task-specific features [38], we divide the feature encoder into the shallow encoder $E_{shallow}(.;\theta_1)$ and the deep encoder $E_{deep}(.;\theta_2)$. We connect the domain predictor $P_{domain}(.;\theta_3)$ after $E_{shallow}(.;\theta_1)$ and connect the emotion label predictor P_{label} after $E_{deep}(.;\theta_2)$. Gradient reversal layer (GRL) is added between $E_{shallow}(.;\theta_1)$ and $P_{domain}(.;\theta_3)$ to realize the antagonistic effect of a two-player game, which forces the shallow encoder to generate domain-independent representation for cross-subject alignment. Therefore, the loss function of cross-subject alignment module is defined as:

$$\mathcal{L}_{cross-subject} = \mathcal{L}_{label} + \alpha \times \mathcal{L}_{domain} \tag{3}$$

where α is the hyperparameter, which adjusts the weights of $\mathcal{L}_{domain}$. The item $\mathcal{L}_{label}$ is the emotion classification loss on the source domain D_S, which is defined as:

$$\mathcal{L}_{label} = H(\hat{y}, y) = -\sum_{m=1}^{M} y_m \log \hat{y}_m \tag{4}$$

where H is the cross-entropy loss, $\hat{y}$=$P_{label}(E_{deep}(E_{shallow}(X;\theta_1);\theta_2);\theta_3)$ is the predicted value of the emotion label predictor P_{label} on source domain data, and y is the ground truth emotion label, and M represents the number of emotion states. Thus, the seeking process of the parameters θ_1 and θ_2 can be represented as :

$$(\hat{\theta_1}, \hat{\theta_2}) = \arg\min_{(\theta_1,\theta_2)} \mathcal{L}_{label} \tag{5}$$

The item $\mathcal{L}_{domain}$ in (3) is the loss of domain classification on both source and target domains (D_S, D_T), which is defined as:

$$\mathcal{L}_{domain} = -(p \cdot log(\hat{p}) + (1-p) \cdot log(1-\hat{p})) \tag{6}$$

where $\hat{p} = P_{domain}(E_{shallow}(X;\theta_1);\theta_3)$ is the predicted value of the domain label predictor $P_{domain}(\theta_3)$, and $p \in \{0, 1\}$ is the real domain label. The case $p = 0$ means the input data comes from source domain D_S, and the other case $p = 1$ means it comes from the target domain D_T. Then, the parameters θ_3 of domain predictor P_{domain} and the parameters θ_1 of shallow encoder $E_{shallow}$ are optimized according to the following expression:

$$\hat{\theta_3} = \arg\min_{\theta_3} \mathcal{L}_{domain} \tag{7}$$

$$\begin{aligned}\hat{\theta_1} &= \arg\max_{\theta_1} \mathcal{L}_{domain} \\ &= -\arg\min_{\theta_1} \mathcal{L}_{domain}\end{aligned} \tag{8}$$

The illustration of the dynamic source domain selection algorithm is shown in Fig. 3. We take the class dissimilarity among multiple source domains into account and adopt a self-paced learning strategy. It dynamically selects the source domain data

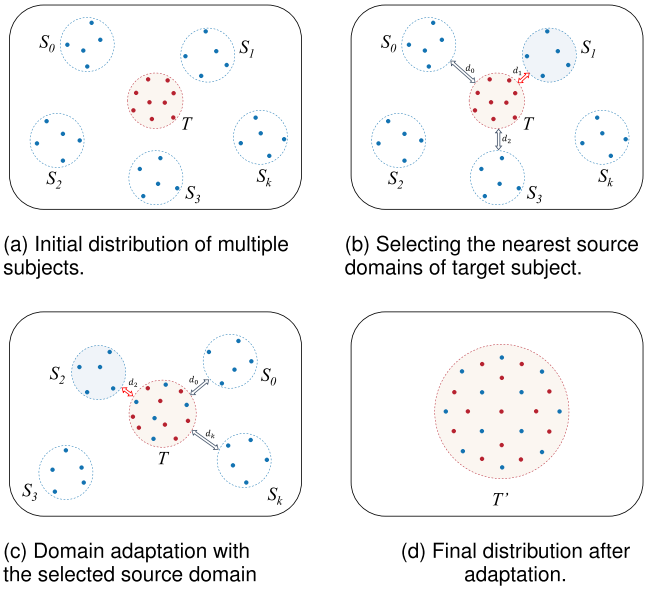


Fig. 3. Illustration of the self-paced learning based source domain selection algorithm.

that are closest to the target domain in the projection space to participate in the training process. This strategy simulates the human cognitive mechanism that processes the learning task from easy to hard. In other words, it gradually transforms the multi-source domain adaptation problem into multiple single-source domain adaptation problems rather than simply treating multiple source domains as a whole single-source domain. Specially, the entire dataset can be defined as $\{S_1, S_2, \ldots, S_N, T\}$, where S_i is the i-th source domain with labels, and T is the target domain without labels. The dynamic source selection algorithm is shown in Algorithm 1. At the beginning of the training stage, the selected domain set $D_{selected}$ only includes the target domain T, that is $D_{selected} = T$. For the remaining source domains $D_{unselected}$, we select the source domain with the minimum distance $dis(S_i, T)$:

$$S_j = \underset{S_i \in D_{unselected}}{\operatorname{argmin}} \; dis\,(S_i, T) \tag{9}$$

$$dis(S_i, T) = -H(P_{domain}(S_i \cup T, \theta_3), p) \tag{10}$$

where $H(a, b) = -[a \cdot \log(b) + (1-a) \cdot \log(1-b)]$ denotes the cross-entropy, $P_{domain}(S_i \cup T, \theta_3)$ is the predicted value of domain predictor and p is the ground truth of domain. The smaller H indicates that the source and target domains are more easily distinguishable. At this point, $-H$ implies a larger distance between them. Compared to unordered random selection, this dynamic selection of training samples can further alleviate the problem of model instability during adversarial training due to the class dissimilarity among multiple source domains.

Cross-subject alignment module was performed for both EEG and eye movement modal data. We define the feature vector of EEG as $X \in \mathbb{R}^{n\times d_1}$, where n denotes the numbers of samples, and d_1 denotes the dimensions of EEG modality in a certain time window, $d_1 = N_C \times N_F$, N_C is the number of EEG channels, and N_F is the number of frequency bands. We define the feature

Algorithm 1: The Proposed Self-Paced Adversarial Domain Adaptation Algorithm.

Input: N source domains $\boldsymbol{S} = \{S_1, S_2, \ldots, S_N\}$, ($S_i = \{x_1, x_2, \ldots, x_n\}$, n is the number of samples) and target domain T
Output: Source domains $\boldsymbol{S'} = \{S'_1, S'_2, \ldots, S'_N\}$ and target domain T' after cross-subject alignment
Initialization: parameters θ_1, θ_2 and θ_3 in cross-subject alignment module, hyperparameters α, β
while epoch <max_epoches **do**
  $D_{unselected} = \boldsymbol{S}$
  for $k = 1$ to N **do**
    $T' = E_{shallow}(T, \theta_1)$
    $\boldsymbol{d} = \varnothing$
    for each $S_i \in D_{unselected}$ **do**
      $S'_i = E_{shallow}(S_i, \theta_1)$
      $d_i = dis(S'_i, T)$
      $\boldsymbol{d} = \boldsymbol{d} \cup \{d_i\}$
    end for
    $d_j = min\ \boldsymbol{d}$
    $\hat{y} = P_{label}(E_{deep}(E_{shallow}(S_j)))$
    $\hat{p} = P_{domain}(E_{shallow}(S_j \cup T))$
    $\mathcal{L}_{label} \leftarrow$ calculated by Eq. (4)
    $\mathcal{L}_{domain} \leftarrow$ calculated by Eq. (6)
    $\mathcal{L}_{cross-subject} = \mathcal{L}_{label} + \alpha \times \mathcal{L}_{domain}$
    update network.
    $D_{unselected} = D_{unselected} - S_j$
  end for
end while
$\boldsymbol{S'} = E_{shallow}(\boldsymbol{S})$
$T' = E_{shallow}(T)$
return $\boldsymbol{S'}$, T'

vector of eye movement modality as $Y \in \mathbb{R}^{n \times d_2}$, where d_2 denotes the number of statistical features of eye movement modality. The output feature vectors of EEG and eye movement modalities after cross-subject alignment can be represented as $X' \in \mathbb{R}^{n \times D_1}$, $Y' \in \mathbb{R}^{n \times D_2}$, respectively.

During the optimization process, the parameters of this module are updated as follows:

$$\theta_1 \leftarrow \theta_1 - lr\left(\frac{\partial \mathcal{L}_{label}}{\partial \theta_1} - \alpha \times \frac{\partial \mathcal{L}_{domain}}{\partial \theta_1}\right) \tag{11}$$

$$\theta_2 \leftarrow \theta_2 - lr\frac{\partial \mathcal{L}_{label}}{\partial \theta_2} \tag{12}$$

$$\theta_3 \leftarrow \theta_3 - lr \times \alpha\frac{\partial \mathcal{L}_{domain}}{\partial \theta_3} \tag{13}$$

where lr is the learning rate.

B. Cross-Modal Alignment

In emotion recognition, multi-modal data can provide richer discriminative features than single-modal data, but at the same time, it also brings greater challenges to recognition due to the heterogeneity between modalities. Inspired by ALBEF [33], in order to capture the latent effective intra-modal information and alleviate the structural heterogeneity between modalities, we set the inter-modal contrastive learning loss on the basis of deep encoders with self-attention mechanism before fusion. For purposes of illustration, we define $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^{n \times d_x}$, $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_n\} \in \mathbb{R}^{n \times d_y}$ as the feature vectors of EEG and eye movement modalities after the projection of shallow encoder $E_{shallow}$ in the cross-subject alignment, where n is the number of samples, and d_x and d_y are the dimensions of feature vectors of two modalities. The self-attention encoder module consists of three self-attention layers. In addition, a residual connection layer is added to alleviate the problem of gradient disappearance and network degradation.

In the self-attention layer, queries (Q), keys (K), and values (V) of both modalities are calculated through different linear projections:

$$K_X = XW_X^K, Q_X = XW_X^Q, V_X = XW_X^V \tag{14}$$

$$K_Y = YW_Y^K, Q_Y = YW_Y^Q, V_Y = YW_Y^V \tag{15}$$

where parameter matrices W^Q, W^K, W^V are updated by the back propagation in the training stage, which are used to generate Q, K and V. The correlation between different features within the single modality is modeled through scaled dot-product. The final feature matrix of two modalities with self-attention is computed as:

$$f(X) = softmax\left(\frac{Q_X K_X^T}{\sqrt{d_x}}\right) V_X \tag{16}$$

$$g(Y) = softmax\left(\frac{Q_Y K_Y^T}{\sqrt{d_y}}\right) V_Y \tag{17}$$

Since the heterogeneity between modalities limits the performance of the model, it is necessary to align the modalities before fusion. Inspired by contrastive representation learning, we construct the contrastive loss between synchronous features and asynchronous features to reduce the structural differences between modalities. Based on the fact that emotional stimuli can be reflected in human physiological activities such as brain electricity, facial expressions, eye movement and etc., the similarity between synchronous modalities is higher than that of asynchronous modalities. Inspired by the idea of MOCO [39], we set the momentum encoders, which share the same structure with the self-attention modules to obtain two queues with recently stored M EEG-eye movement sample pairs. The feature vectors after the projection of momentum encoders are denoted as $f'(X)$ and $g'(Y)$. The similarity between different modal representations is defined as:

$$s(X, Y) = f(X)^T g'(Y) \tag{18}$$

$$s(Y, X) = g(Y)^T f'(X) \tag{19}$$

For each sample that contains two modalities, the pairwise modal similarity $p^x(X) = \{p_1^x(X), p_2^x(X), \ldots, p_M^x(X)\}$ and $p^y(Y) = \{p_1^y(Y), p_2^y(Y), \ldots, p_M^y Y\}$ with the representations stored in the momentum queue are calculated separately as (20)

and (21).

$$p_m^x(X) = \frac{\exp(s(X, Y_m)/\tau)}{\sum_{i=1}^{M} \exp(s(X, Y_i)/\tau)} \tag{20}$$

$$p_m^y(Y) = \frac{\exp(s(Y, X_m)/\tau)}{\sum_{i=1}^{M} \exp(s(Y, X_i)/\tau)} \tag{21}$$

where τ is a temperature parameter, X_m and Y_m represent the m-th EEG modality representation and the m-th eye movement modality representation stored in the momentum queue, respectively.

The ground-truth similarity of different modalities are represented as $q^x(X)$ and $q^y(Y)$. Specifically, $q^x(X) = \{q_1^x(X), q_2^x(X), \ldots, q_M^x(X)\}$ denotes the similarity between EEG modality representation extracted by the self-attention module and the eye movement features preserved in the momentum queue. Likewise, $q^y(Y) = \{q_1^y(Y), q_2^y(Y), \ldots, q_M^y(Y)\}$ denotes the similarity between the eye movement modality extracted by the self-attention module representation and the EEG modality representation preserved in the momentum queue. The probability of synchronized EEG-eye movement pairs is 1; Otherwise, it is set to 0. To be specific, for sample X_j,

$$q_i^x(X_j) = \begin{cases} 1, t_{X_j} = t_{Y_i} \\ 0, t_{X_j} \neq t_{Y_i} \end{cases} \tag{22}$$

For sample Y_j,

$$q_i^y(Y_j) = \begin{cases} 1, t_{Y_j} = t_{X_i} \\ 0, t_{Y_j} \neq t_{X_i} \end{cases} \tag{23}$$

The modal contrast loss is defined as the cross-entropy H between p and q:

$$\mathcal{L}_{contrast} = \frac{1}{2}\mathbb{E}_{(X,Y)\sim D}[H(q^x(\boldsymbol{X}), p^x(\boldsymbol{X})) + H(q^y(\boldsymbol{Y}), p^y(\boldsymbol{Y}))] \tag{24}$$

C. Multi-Modal Fusion

Although EEG and eye movement data are different in sensing mode, they are semantically consistent because neural patterns of the brain and human behavior interact with each other in response to the same emotional stimulus. In order to capture the higher-order semantic information between modalities and improve the reliability of emotion recognition, we design the modal matching loss based on contrastive learning to constrain the cross-attention encoder to generate latent inter-modal information. Specifically, the representation A and B mapped from cross-modal alignment modules will be concatenated, and they are fed to the multi-modal cross-attention encoder as the input $Z = [A \quad B]$ for further modal fusion at the feature level, where $Z \in \mathbb{R}^{n\times(d_x+d_y)}$. The expressions of query, key, and value are as follows:

$$K_Z = ZW_Z^K = [A \quad B]\begin{bmatrix} W_A^K \\ W_B^K \end{bmatrix} = K_A + K_B \tag{25}$$

$$Q_Z = ZW_Z^Q = [A \quad B]\begin{bmatrix} W_A^Q \\ W_B^Q \end{bmatrix} = Q_A + Q_B \tag{26}$$

$$V_Z = ZW_Z^V = [A \quad B]\begin{bmatrix} W_A^V \\ W_B^V \end{bmatrix} = V_A + V_B \tag{27}$$

$$Attention(Q_Z, K_Z, V_Z) = softmax\left(\frac{Q_Z K_Z^T}{\sqrt{d_x + d_y}}\right) V_Z \tag{28}$$

where $K_A = AW_A^K$, $Q_A = AW_A^Q$, $V_A = AW_A^V$, and $K_B = BW_B^K$, $Q_B = BW_B^Q$, $V_B = BW_B^V$. The item $Q_Z K_Z^T V_Z$ in cross attention of the multi-modal fusion module is:

$$\begin{aligned} Q_Z K_Z^T V_Z &= (Q_A + Q_B)(K_A + K_B)^T (V_A + V_B) \\ &= Q_A K_A^T V_A + Q_B K_A^T V_A + Q_A K_B^T V_A + Q_B K_B^T V_A \\ &\quad + Q_A K_A^T V_B + Q_B K_A^T V_B + Q_A K_B^T V_B + Q_B K_B^T V_B \end{aligned} \tag{29}$$

As can be seen from the above formula, the cross-attention module considers both intra-modal and inter-modal information. These representations are followed by a fully connected layer and softmax layer for binary classification, and the predicted probability of them is defined as u^{xy}. In the multi-modal cross-attention module, we utilize contrastive learning to capture the semantic similarity between modalities. Synchronously collected EEG signals and eye movement data constitute a positive sample pair, i.e (X_j, Y_j^+), while asynchronous modal data constitute a negative sample pair, i.e. (X_i, Y_j^-). For arbitrary sample pair (X_i, Y_j), we expect that it could be closer to the positive samples that are composed of synchronous modalities and be far away from the negative samples that are composed of asynchronous modalities. The modal matching label v^{xy} of the positive sample will be set to 1; Otherwise, it will be set to 0. The modal matching loss is represented as follows:

$$\mathcal{L}_{match} = E_{(X,Y)\sim D} H(v^{xy}, u^{xy}) \tag{30}$$

In order to ensure the discriminability of the model on the emotion classification, we also use the classification loss on the positive sample pairs:

$$\mathcal{L}_{class} = E_{(X,Y)\sim D} H(y^{class}, p^{class}) \tag{31}$$

where y_{class} is the ground-truth emotion label, p^{class} is the predicted value. Thus, the overall loss of cross-modal alignment module and multi-modal fusion module is :

$$\mathcal{L}_{cross-modal} = \mathcal{L}_{contrast} + \mathcal{L}_{match} + \mathcal{L}_{class} \tag{32}$$

Considering the problem of missing modality may cause the failure of the model in the real world, we make full use of the discriminative information in the cross-subject alignment module at the decision level. In the cross-subject alignment module, deep encoders tend to further generate class discriminative information based on the subject-independent representations, and the emotion label predictor can be generalized from the source domain to the target domain. The output P_{fusion} of the

TABLE I
THE SUMMARY OF DATASETS

	SEED		SEED-IV		SEED-V	
	EEG	Eye	EEG	Eye	EEG	Eye
Features/sample	310	33	310	31	310	33
Num. subjects	12		15		16	
Num. trials	540		1080		720	
Num. samples	30312		37425		29168	

The window size of each sample is set as 4 seconds. Features/sample means the final extracted feature dimensions of a sample.

final layer of our model can be expressed as:

$$P_{fusion} = softmax(p_{eeg} + p_{eye} + p_{cross-modal}) \quad (33)$$

where p_{eeg} and p_{eye} are the normalized vectors before the softmax layer in the label predictor of cross-subject modules, respectively. The index of the maximum value of the output P_{fusion} is the predicted value.

IV. EXPERIMENTS

In this section, we conduct extensive experiments of our proposed method and other comparison methods on three multi-modal emotional datasets.

A. Data Materials

The SEED, SEED-IV and SEED-V datasets with EEG and eye movement modalities were provided by the BCMI laboratory. The summary description of datasets is shown in Table I. The SEED dataset contains 540 multi-modal trials with positive/negative/disgust/neutral/fear emotion states from 12 subjects. Each of these subjects participated in a total of 3 experimental sessions, with 15 trials in each session. The SEED-IV dataset contains 1080 multi-modal trials with happy/sad/neutral/fear emotion states from 15 subjects. Each of these subjects participated in a total of 3 experimental sessions, with 24 trials in each session. The SEED-V dataset contains multi-modal 720 trials with happy/sad/neutral/fear emotion states from 16 subjects. Each of these subjects participated in a total of 3 experimental sessions, with 15 trials in each session. The EEG signals were collected with the 62-channel ESI NeuroScan System and the eye movement modality was collected with SMI eye-tracking glasses. For the multi-modal setting, the sliding windows of both EEG and eye movement modalities are set as 4 seconds. We directly use the features that have been processed by authorities in the experiments. The raw EEG signals are filtered with a 0.2–50 Hz bandpass filter, and eye blinking artifacts are removed with a threshold algorithm. Finally, the processed EEG signals are downsampled from 1000 Hz to 200 Hz. According to [40], [41], differential entropy features (DE) are extracted from five frequency bands of EEG modality. The linear dynamic system approach is applied to smooth DE features [42]. In addition, the eye movement modality data contains statistical information such as pupil diameter and blink duration, and computational statistics such as temporal and frequency features.

B. Experiment Settings

We conduct experiments to verify the superiority of our proposed model on the task of cross-subject multi-modal emotion recognition. In addition, we also adopted two common tasks, namely intra-subject multi-modal and cross-subject single-modal emotion recognition to verify our cross-subject alignment module and multi-mode fusion module, respectively.

1) Cross-subject multi-modal setting: In this setting, a leave-one-subject-out cross-validation scheme is adopted to verify our overall model. For each subject, we use his/her samples with EEG and eye movement modalities as target domain data, and the multi-modal data of the remaining subjects as source domain data. The representation of the source domain after cross-subject alignment is used as the training data set of the downstream multi-modal fusion network, while the cross-subject aligned representation of the target subject is used as the test data set. During the entire training process, the label data of the source domain is not visible.

2) Intra-subject multi-modal setting: For each subject, the data of first 9 trials of SEED dataset/ first 14 trials of SEED-IV dataset/ first 9 trials of SEED-V dataset is divided into the training set, and the remaining data is divided into the test set. In each experiment, our proposed multi-modal alignment and fusion module and other compared methods are performed on one subject's EEG signals and eye movement data. The average results of all subjects are recorded.

3) Cross-subject unimodal setting: In this setting, we adopted the same leave-one-subject method as the cross-subject multi-modal setting. We utilize the single EEG modality data of source and target domains to perform our proposed self-paced adversarial domain adaptation model. For each subject, we use his/her EEG data from all trials as target samples, and the EEG data of the rest subjects as source samples.

C. Performance Metrics

For emotion classification, accuracy and standard deviation are the most common metrics to measure the performance of the model. We also introduce Macro-F1 score to comprehensively measure the performance of models in multi-classification scenarios. The calculation formula of Macro-F1 in n-class problem is as follows:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (34)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (35)$$

$$Precision_{macro} = \frac{\sum_{i=1}^{n} Precision_i}{n} \quad (36)$$

$$Recall_{macro} = \frac{\sum_{i=1}^{n} Recall_i}{n} \quad (37)$$

$$Macro{-}F1 = \frac{2 * Precision_{macro} * Recall_{macro}}{Precision_{macro} + Recall_{macro}} \quad (38)$$

In addition, AUC metric, calculated by the area under the ROC curve is also introduced to evaluate the discriminability of models.

D. Comparison Methods

The comparison emotion recognition methods of our proposed method are selected from two aspects, i.e. multi-modal fusion based method, and cross-subject alignment based methods. However, among the few cross-subject multi-modal emotion recognition methods, some methods such as SDAE [43], did not strictly perform domain alignment, simply using source-only (SO) experimental setting instead of making better improvements for domain drift. We also compare with some recent cross-subject multi-modal emotion recognition methods, such as CFDA-CSF [35], CMSLNet [36] and CiABL [37]. In addition, to be fair, we compare our multi-modal fusion module with other multi-modal based fusion methods in the intra-subject multi-modal setting, and compare our cross-subject alignment module with other cross-subject alignment based methods and source-only methods in the cross-subject unimodal setting. Finally, in the cross-subject multi-modal setting, SO method and five effective cross-subject based methods are selected to combine with three different multi-modal fusion methods as the comparison methods. Notably, we compared our method with some of these comparison methods, including CFDA-CSF, CMSLNet, JAGP, FSA-TSP and MS-ADA, using experimental results provided by the authors in their original papers. To ensure fair comparisons, all methods used the same data processing and experimental settings. The selected multi-modal fusion based methods and cross-subject alignment based methods are as follows:

1) Multi-modal fusion based methods, including MKL [28], CCA based methods, such as CCA [25], KCCA [26], DCCA [27], DCCA-AM method [9], deep learning methods such as BDAE [11], TMVC [44] and MDDF [45].

2) Cross-subject alignment based methods, including conventional transfer learning methods, such as TCA [14], KPCA [46], TPT [15], deep domain confusion method using maximum mean discrepancy (MMD) [16], adversarial domain adaptation networks such as DANN [17], MADA [18], ADDA [19], MDDA [20], MDAN [21], JAGP [22], FSA-TSP [23] and MS-ADA [24].

E. Parameter Settings

In the cross-subject alignment module, there were two shallow layers with 256 and 128 hidden units and one deep layer with 32 hidden units in the encoder. The output layer of the label classifier has three units corresponding to three emotions, and the output layer of the domain predictor has two units corresponding to the source domain and target domain. In the cross-modal alignment module, two hidden layers with a multi-head attention mechanism are set on both EEG and eye movement modalities. The structure for the EEG branch is 310-128-64 with two attention heads, and the structure for the eye movement branch is 33/31-128-64 with one attention head. ReLU is used as the activation function. In addition, a dropout layer and a norm layer are adopted to avoid overfitting and internal covariate shifts. The structure of the cross-attention module is 128-64-32. The grid search is adopted to select the optimal parameters. The queue size searches from sets $\{256, 512, 1024, 2048, 4096\}$, the momentum parameter searches from sets $\{0.5, 0.7, 0.8, 0.85, 0.9, 0.95\}$ and the temperature parameter t searches from sets $\{0.2, 0.4, 0.5, 0.6, 0.8\}$. We used the AdamW optimizer, and the training epochs are searched from 100 to 1000 with a step size of 50, and the learning rate is within the range of $\{1e-4, 1e-3, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$.

V. Results and Analysis

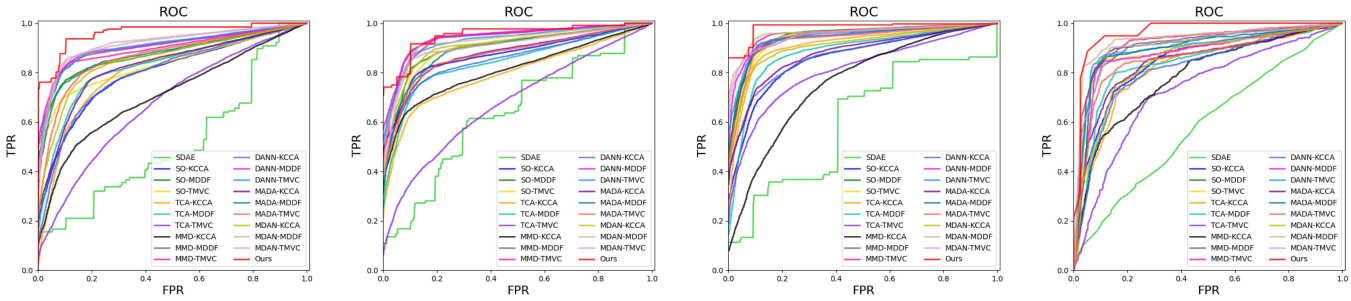
A. Classification Performance

Performance of our proposed cross-subject multi-modal emotion recognition method and the comparison algorithms on three datasets are illustrated in Table II. The experimental results show the effectiveness and stability of our proposed method are better than other algorithms. As can be seen in Table II, the average accuracy of our proposed method on the task of cross-subject multi-modal emotion recognition reach 94.96%, 89.82% and 89.22%, respectively, on SEED, SEED-IV and SEED-V datasets, which are obviously better than other comparison methods. We plot the ROC curves of our method and other comparison algorithms on SEED dataset in Fig. 4. It can be seen that the ROC curve of our method is significantly closer to the upper left corner than other algorithms, and our method has a higher AUC. Fig. 5 illustrates the confusion matrices of the performance of single modality under the action of our cross-subject alignment module, and the performance of our overall cross-subject multi-modal fusion emotion recognition model. As can be seen from Fig. 5, it is obvious that the recognition performance of multi-modal fusion is better than that of single modality, and the reason may be that our multi-mode fusion module can mine higher-order semantic information.

In addition, the experiment results of intra-subject multi-modal setting and cross-subject unimodal setting are illustrated in Tables III and IV. Figs. 6 and 7 show the overall performance comparisons of cross-subject alignment methods and multi-modal fusion methods, respectively. It is worth noting that even if we remove one of the cross-subject alignment module and multi-modal fusion module from our method, it still has better classification performance compared with other algorithms. In Table III, the average accuracy of our cross-modal alignment and multi-modal fusion modules on the task of intra-subject multi-modal emotion recognition reaches 91.36%, 89.56% and 80.10%, respectively, on SEED, SEED-IV and SEED-V datasets. This demonstrates that our multi-modal fusion strategy based on attention mechanisms and contrastive representation learning can mine latent higher-order emotion-related semantic information, which assists the model in making more accurate and reliable emotion predictions. In Table IV, the average accuracy of our cross-subject alignment module on the task of cross-subject single-modal emotion recognition reaches 92.41%, 85.60% and 86.29%, respectively, on SEED, SEED-IV and datasets. It is clear that our self-paced adversarial domain

TABLE II
THE EXPERIMENT RESULTS OF DIFFERENT METHODS ON THE TASK OF CROSS-SUBJECT MULTI-MODAL EMOTION RECOGNITION (%)

Cross-subject Alignment	Multi-modal Fusion	SEED			SEED-IV			SEED-V		
		Acc	Macro-F1	AUC	Acc	Macro-F1	AUC	Acc	Macro-F1	AUC
SDAE		61.13± 6.99	60.71± 7.03	70.78± 5.28	64.25± 10.75	65.89± 11.15	72.18± 9.12	58.25± 7.55	50.67± 9.49	70.69± 4.86
SO	KCCA	73.12±13.27	74.55±13.66	79.64± 9.95	61.74± 6.53	60.15± 8.33	70.61± 4.77	57.34± 7.92	49.71± 9.72	70.14± 5.07
	TMVC	85.33± 9.83	85.97± 9.88	88.94± 7.42	78.98±10.01	80.03± 9.98	85.81± 6.37	79.97±11.29	80.42±11.68	87.01± 7.38
	MDDF	87.88± 8.75	76.58± 4.37	90.86±6.60	78.44±10.66	79.33±11.31	85.14± 7.26	78.93±13.83	79.72±15.74	86.41± 8.94
TCA	KCCA	74.39±10.22	75.22±10.21	80.54± 7.74	60.18±10.24	61.77± 9.13	68.68± 7.63	40.81±12.30	36.01±14.04	61.47± 7.87
	TMVC	62.83±12.35	61.66± 9.63	72.05± 9.22	68.89± 8.04	69.73± 8.25	78.52± 5.57	68.30± 9.33	66.90±10.88	78.70± 5.82
	MDDF	80.14±10.50	80.76±10.44	85.03± 7.89	66.72± 8.69	66.81±10.16	77.00± 5.91	64.86± 8.48	60.38±10.49	76.22± 5.58
MMD	KCCA	72.33±12.63	73.08±18.85	78.90± 9.61	68.53± 9.05	66.87± 8.29	75.48± 6.38	61.12± 8.92	57.22±13.55	73.09± 6.56
	TMVC	90.16± 5.92	90.78± 5.65	92.43± 4.47	80.70± 8.14	81.68± 7.70	86.76± 5.56	81.43±11.93	81.77±13.03	88.05± 7.58
	MDDF	90.57± 5.34	91.25± 4.97	92.90± 4.02	83.96± 9.14	84.94± 8.63	89.10± 6.29	81.54±11.97	81.86±13.06	88.09± 7.59
DANN	KCCA	76.77±11.39	77.95±10.96	82.39± 8.62	76.14± 8.55	76.72± 8.43	82.30± 6.29	78.38±10.02	77.00±11.55	85.09± 6.43
	TMVC	89.86± 7.89	90.37± 7.52	92.20± 5.90	84.52± 6.78	84.63±6.89	88.96± 4.91	86.36± 8.15	86.77±7.94	90.80± 5.26
	MDDF	90.92± 8.22	91.30± 8.11	93.18± 6.14	83.45± 7.06	84.04± 6.65	88.11± 5.04	86.75± 9.98	87.55± 9.48	91.15± 6.66
MADA	KCCA	77.08± 8.71	78.62± 8.88	82.58± 6.57	68.16± 8.21	67.75± 9.93	75.77± 6.46	46.70±20.80	38.27±26.03	63.86±14.24
	TMVC	85.05± 8.59	85.79± 8.25	88.61± 6.47	72.85± 8.51	74.18± 9.13	81.92± 5.33	81.80±11.84	80.93±13.41	88.06± 7.51
	MDDF	86.72± 8.82	87.69± 8.01	89.98± 6.63	73.09± 9.32	74.53± 9.26	81.67± 6.35	85.76±11.69	85.53±12.46	90.54± 7.58
MDAN	KCCA	77.65±10.61	78.59±10.72	83.03± 7.94	76.17± 6.62	76.50± 7.56	82.28± 5.04	74.68±11.45	75.55±12.19	83.25± 7.39
	TMVC	85.91± 7.38	86.19± 7.44	89.39± 5.54	77.86± 7.88	78.21± 8.23	84.63± 5.63	81.60±10.70	82.21±10.43	88.19±6.85
	MDDF	91.40± 6.44	91.92± 5.87	93.53± 4.86	79.18± 7.14	79.51± 7.45	85.34± 4.94	80.97± 9.60	81.74±10.75	87.68± 6.31
CFDA-CSF		90.04±5.46	-	-	89.60± 6.65	-	-	80.01±11.30	-	-
CMSLNet		-	-	-	83.15±9.84	-	-	87.32±14.81	-	-
CiABL		64.79±12.86	66.81±12.22	75.03±10.88	61.00±12.84	61.39±13.35	74.02±8.72	73.19± 7.55	74.34±8.13	83.40±4.70
Ours		**94.96± 5.27**	**95.21± 7.96**	**96.20± 3.98**	**89.82± 6.22**	**90.03± 6.19**	**93.01± 4.25**	**89.22± 9.59**	**89.95± 8.85**	**93.11± 5.97**



(a) Negative VS. Neutral&Positive (b) Neutral VS. Negative&Positive (c) Positive VS. Negative&Neutral (d) Macro Average

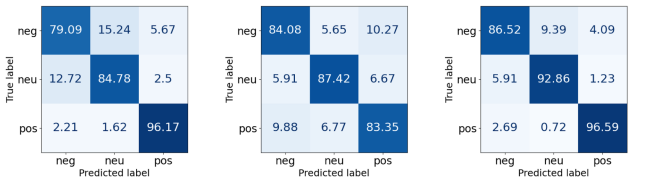
Fig. 4. ROC curves of the proposed method and comparison algorithms on the SEED dataset.

TABLE III
THE EXPERIMENT RESULTS OF DIFFERENT METHODS ON THE TASK OF INTRA-SUBJECT MULTI-MODAL EMOTION RECOGNITION (%)

Methods	SEED			SEED-IV			SEED-V		
	Acc	Macro-F1	AUC	Acc	Macro-F1	AUC	Acc	Macro-F1	AUC
MKL	74.71± 14.76	75.68±16.25	80.25±11.03	74.62±15.00	68.70±28.61	83.35±9.59	59.48±13.27	55.27±13.80	74.04± 7.50
CCA	68.49± 18.38	73.51±20.65	75.72±13.82	67.93±14.01	69.05±13.99	77.33± 9.02	56.57±14.76	52.28±16.14	73.09± 8.63
KCCA	74.67±16.20	77.93±16.85	80.31±12.17	68.62± 9.41	70.57±10.94	78.42± 7.61	53.62±13.19	51.34±14.21	69.71±8.78
DCCA	75.36± 9.85	78.38± 8.87	81.01± 7.51	72.89±13.33	73.85±10.11	81.75±9.23	59.99±10.51	53.32±14.60	73.75± 7.43
DCCA-AM	90.27± 9.02	91.18±10.53	92.54± 6.91	83.12± 9.36	83.93±12.20	89.82± 5.95	73.34±12.38	70.30±17.00	83.67± 8.68
BDAE	88.31±10.42	88.71± 8.87	91.07± 7.97	75.89±15.51	74.25±17.61	84.64±10.13	63.62±12.76	57.93±17.74	77.43± 8.56
TMVC	90.35± 7.43	92.45± 6.58	92.60± 5.71	85.26±12.21	85.74±13.42	90.64± 7.89	70.09±13.85	70.89±14.40	84.04± 7.54
MDDF	90.43±10.28	91.35± 9.84	92.61± 7.96	84.02±11.95	83.11±15.46	89.89± 8.02	70.89±13.33	71.81±13.69	84.64± 6.88
Ours	**91.36± 7.73**	**92.25± 6.81**	**93.10± 6.16**	**89.56±11.54**	**89.63±10.15**	**93.96± 6.31**	**80.10±11.34**	**76.89±14.12**	**87.84± 7.02**

TABLE IV
THE EXPERIMENT RESULTS OF DIFFERENT METHODS ON THE TASK OF CROSS-SUBJECT UNIMODAL EMOTION RECOGNITION (%)

Methods	SEED			SEED-IV			SEED-V		
	Acc	Macro-F1	AUC	Acc	Macro-F1	AUC	Acc	Macro-F1	AUC
SO	67.63±10.09	67.74± 9.69	75.67±7.55	64.36±12.37	56.34± 6.42	75.89± 8.23	60.98±13.21	60.16±15.16	75.48± 8.48
TCA	74.59± 9.82	74.98±10.31	80.63± 7.49	67.58± 8.62	67.64± 8.89	77.86± 5.64	60.31± 8.71	58.12±10.81	73.65± 5.17
KPCA	76.10±11.16	77.45±10.55	81.97± 8.37	66.94± 9.04	66.78± 9.11	77.25± 5.90	57.16±13.87	56.94±13.01	69.15± 8.57
TPT	71.51±10.71	72.69±11.10	79.06± 8.00	67.85± 9.32	67.09±19.47	79.46± 6.68	39.84 ±9.25	44.64±11.44	66.43±14.04
MMD	86.96± 8.36	87.97± 8.89	90.20± 6.29	77.02± 7.20	78.28± 6.90	84.15± 4.90	64.56±12.64	66.38±13.08	78.22± 8.11
DANN	88.23± 7.97	88.59± 7.85	91.13± 5.99	84.24± 7.44	84.63± 6.98	89.13± 5.05	85.90±10.42	86.61±10.03	90.86± 6.42
MADA	83.32± 8.59	84.00± 9.07	87.42± 6.49	72.34± 9.11	73.51± 8.78	81.15± 5.66	75.26± 9.81	73.92±11.50	83.69± 6.29
ADDA	65.87± 6.32	65.75± 6.85	74.37± 4.69	63.07± 8.28	63.05± 8.79	72.22± 6.24	40.81± 8.66	33.29±10.81	61.47± 5.40
MDDA	61.13± 7.84	61.07± 8.00	70.76± 5.90	65.08± 8.66	67.88± 7.72	76.40± 5.78	45.70±11.35	41.13±13.99	63.89± 6.62
MDAN	82.47±10.38	83.31±10.36	86.79±7.81	72.48± 7.78	74.10± 8.20	81.51± 5.07	72.78±12.49	73.45±12.11	82.89± 7.67
JAGP	-	-	-	78.77± 8.15	-	-	75.43±10.72	-	-
FSA-TSP	91.79± 7.50	91.76± 7.54	93.47± 6.03	85.45± 9.81	85.33± 9.22	89.77± 6.70	-	-	-
MS-ADA	86.16± 7.87	-	-	59.29±13.65	-	-	-	-	-
Ours	**92.41± 7.43**	**92.54±7.18**	**94.26± 5.23**	**85.60± 7.49**	**85.72± 7.67**	**90.10± 5.12**	**86.29± 9.80**	**86.83±9.24**	**91.16± 6.16**



(a) EEG modality (b) Eye modality (c) Multi-modal fusion

Fig. 5. The confusion matrices of our proposed method on the task of cross-subject emotion recognition with different modalities on the SEED dataset. (a) shows the performance of EEG modality under the cross-subject alignment module, (b) shows the performance of eye movement modality under the cross-subject alignment module, and (c) shows the performance of our overall cross-subject multi-modal fusion emotion recognition model.

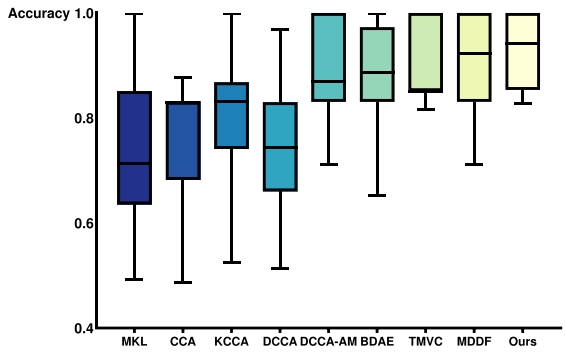


Fig. 6. Performance comparison of multi-modal fusion methods in the intra-subject multi-modal setting on the SEED dataset.

TABLE V
THE ABLATION STUDY RESULTS OF OUR SELF-PACED DOMAIN ADAPTATION METHOD (%)

Cross-subject Alignment	Self-paced Domain Selection	SEED	SEED-IV	SEED-V
		Acc	Acc	Acc
×	×	67.63 ±10.09	64.36±12.37	60.98± 13.21
√	×	88.30 ± 9.09	84.24± 7.44	85.90± 10.42
√	√	**92.41± 7.43**	**85.60± 7.35**	**86.29± 9.80**

TABLE VI
THE ABLATION STUDY RESULTS OF OUR MULTI-MODAL FUSION NETWORK.(%)

Attention Mechanism	$\mathcal{L}_{contrast}$	$\mathcal{L}_{match}$	SEED	SEED-IV	SEED-V
			Acc	Acc	Acc
×	×	×	81.60 ±12.24	77.40±12.02	76.27±11.90
×	×	√	83.31 ± 9.02	78.88± 8.79	79.55± 9.10
×	√	×	84.91 ±10.14	79.18±10.24	80.48± 9.05
×	√	√	85.14 ±11.25	80.26±12.30	80.76±11.60
√	×	×	84.79 ±10.39	80.11±11.25	84.72±10.17
√	×	√	88.18 ± 8.67	82.37± 8.07	85.04± 9.55
√	√	×	88.08 ± 7.67	84.62± 8.24	86.33± 8.79
√	√	√	**94.96 ± 5.27**	**88.89± 6.40**	**89.22± 9.59**

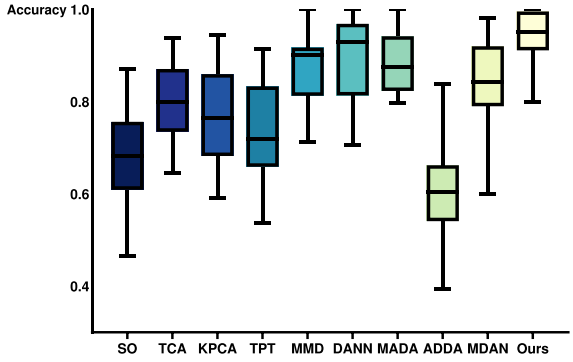


Fig. 7. Performance comparison of cross-subject alignment methods in the cross-subject unimodal setting on the SEED dataset.

adaptation method in the cross-subject alignment module works compared to the method without cross-subject distribution alignment. Our method can effectively extract domain-invariant and class-discriminative features to achieve accurate emotion classification.

B. Ablation Study

1) Validity of Self-Paced Domain Adaptation Network: To obtain subject-independent feature representation to improve the

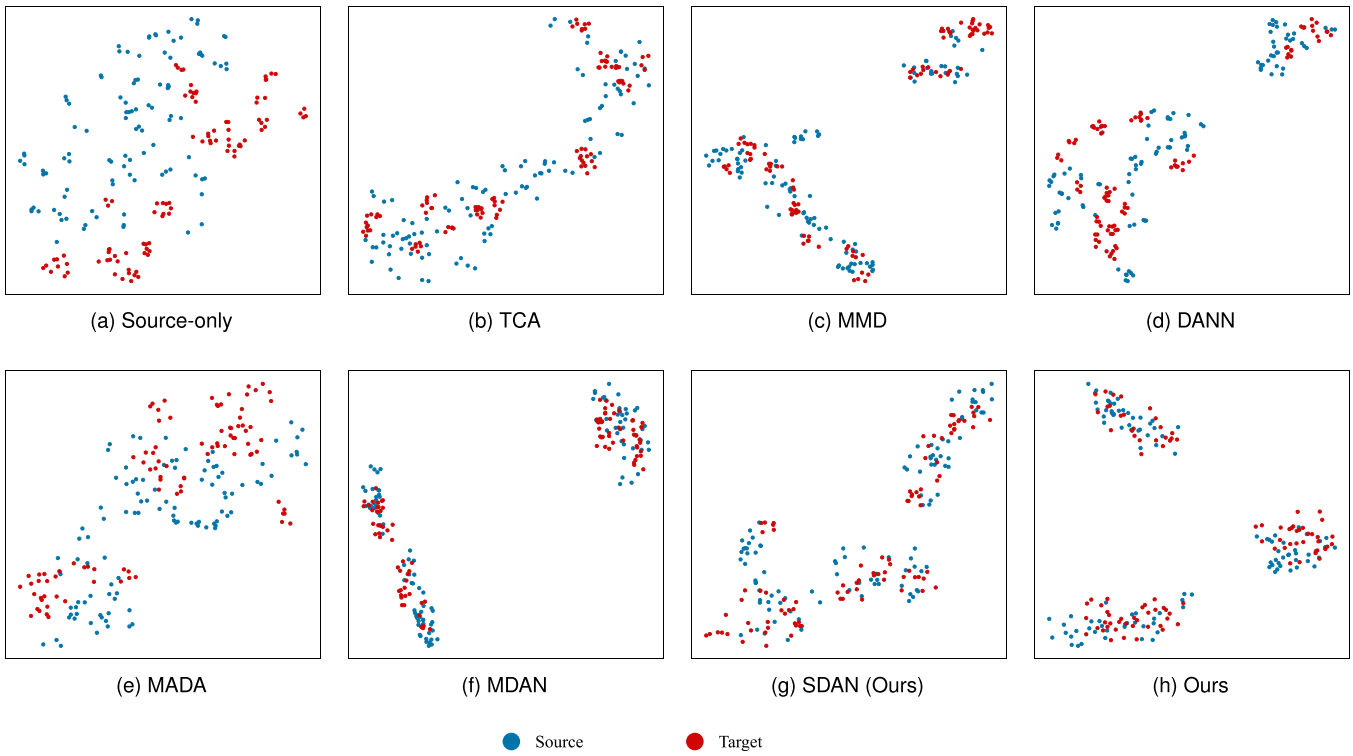


Fig. 8. The T-SNE visualization of different cross-subject emotion recognition methods on the SEED dataset. (a) shows the data distribution of source-only method without cross-subject alignment. (b)-(f) are the distributions by different cross-subject alignment methods. (g) shows the distribution generated by our proposed self-paced adversarial domain adaptation network (SDAN) module, and (h) is the representation space of our proposed cross-subject multi-modal method.

generalization ability of the model, we perform cross-subject alignment in (11), (12), (13). In this section, we investigate the effectiveness of the adversarial domain adaptation network modified by a self-paced domain selection algorithm to improve the performance of cross-subject emotion recognition. Under the same single-modality cross-subject experimental setting, we respectively compare the performance of three methods, that is source-only method without alignment, adversarial domain adaptation alignment without self-paced source domain selection, and adversarial domain adaptation alignment with self-paced source domain selection. The results are shown in Table V. Obviously, the result of our method is better than others, indicating the necessity of cross-subject alignment and the ability of our self-paced adversarial domain adaptation network to extract domain-invariant representation while preserving class discriminative information.

2) Validity of Multi-Modal Fusion Network: Our multi-modal fusion network can extract intra-modal and inter-modal high-order information related to emotional patterns. The attention mechanism is applied to capture potential higher-order semantic representations in both the unimodal branch and the cross-modal branch. The modal contrastive loss is designed to optimize self-attention modules of two modalities to produce modal-aligned representation. In addition, the modal matching loss between paired samples and unpaired samples promotes the network to mine semantic similarity between different modalities. To verify the effectiveness of multi-modal fusion network, we conducted ablation studies on different combinations of attention mechanism, modal contrastive loss, and modal matching loss, and the results are shown in Table VI. It is obvious that the performance can be improved when any of the attention mechanism, modal contrastive loss, and modal matching loss is individually added to the original model.

C. Visualization

In order to verify the domain-invariant discriminative higher-order semantic representation generated by our proposed method, we visualize the features learned by different cross-subject recognition methods in Fig. 8. The representation vectors learned from different methods are mapped into two-dimensional space through the T-SNE method for visualization. Compared with other cross-subject alignment methods, the representation space of the source-only method shows lower class discriminability and the data distribution distance between source domains and target domain is more obvious, as shown in Fig. 8(a). As shown in Fig. 8(b)–(f), it is difficult to distinguish three emotion states in the feature representation space of other comparison methods. It can be seen in Fig. 8(g), our proposed self-paced adversarial domain adaptation network (SDAN) on the single EEG modality can extract domain-invariant features and retain the discriminability of emotion states, but there is still a slight stickiness in negative and neutral emotions. Fig. 8(h) is the final representation space projected by our proposed cross-subject multi-modal emotion recognition method. Compared to SDAN module in Fig. 8(g), samples with the same emotional state are clustered together, and samples with different emotional states show more significant

differentiation in Fig. 8(h). It illustrates that our multi-modal fusion network based on attention mechanism and contrastive learning can effectively extract the higher-order semantic information inherent between the modalities to improve the class-discriminability of embedding representation.

VI. Conclusion

In this paper, we propose the cross-subject multi-modal emotion recognition model, which effectively fuses the complementary information between EEG and eye movement, while alleviating distribution drift among different subjects. First, the self-paced adversarial domain adaptation network is designed to capture domain-invariant features from EEG and eye movement modalities. Second, an attention mechanism is introduced to both cross-modal alignment and multi-modal fusion modules for improving the intra-modal and inter-modal embedding representations. The contrastive learning in both feature and sample levels is applied to constrain feature encoders to extract emotion-related higher-order semantic information. Finally, we used the output of the softmax layer as the predicted value. Experimental results on several multi-modal emotion datasets demonstrate that our method outperforms other state-of-the-art emotion recognition methods. In addition, the visualization illustrates our proposed model can effectively extract the inherent correlation among multiple modalities and generate domain-invariant and class-discriminative representations for accurate multi-modal emotion classification.

References

[1] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affective Comput.*, vol. 10, no. 3, pp. 374–393, Third Quarter 2019.

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

[3] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, 2020, Art. no. 592.

[4] X. Zheng, W. Chen, M. Li, T. Zhang, Y. You, and Y. Jiang, "Decoding human brain activity with deep learning," *Biomed. Signal Process. Control*, vol. 56, 2020, Art. no. 101730.

[5] X. Xing, Z. Li, T. Xu, L. Shu, B. Hu, and X. Xu, "SAE LSTM: A new framework for emotion recognition from multi-channel EEG," *Front. Neurorobot.*, vol. 13, 2019, Art. no. 37.

[6] A. Garg, A. Kapoor, A. K. Bedi, and R. K. Sunkaria, "Merged LSTM model for emotion classification using EEG signals," in *Proc. Int. Conf. Data Sci. Eng.*, 2019, pp. 139–143.

[7] X. Huang et al., "Multi-modal emotion analysis from facial expressions and electroencephalogram," *Comput. Vis. Image Understanding*, vol. 147, pp. 114–124, 2016.

[8] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 1170–1176.

[9] W. Liu, W.-L. Zheng, Z. Li, S.-Y. Wu, L. Gan, and B.-L. Lu, "Identifying similarities and differences in emotion recognition with EEG and eye movements among Chinese, German, and French people," *J. Neural Eng.*, vol. 19, no. 2, 2022, Art. no. 026012.

[10] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process.*, Springer, 2016, pp. 521–529.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

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

[13] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition," *IEEE Trans. Affective Comput.*, vol. 14, no. 3, pp. 2496–2511, Third Quarter 2023.

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

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

[16] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 25, pp. 723–773, 2012.

[17] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, Jan. 2016.

[18] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3934–3941.

[19] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2962–2971.

[20] S. Zhao et al., "Multi-source distilling domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12975–12983.

[21] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8568–8579.

[22] Y. Peng, W. Wang, W. Kong, F. Nie, B.-L. Lu, and A. Cichocki, "Joint feature adaptation and graph adaptive label propagation for cross-subject emotion recognition from EEG signals," *IEEE Trans. Affective Comput.*, vol. 13, no. 4, pp. 1941–1958, Fourth Quarter 2022.

[23] M. Jiménez-Guarneros and G. Fuentes-Pineda, "Cross-subject EEG-based emotion recognition via semisupervised multisource joint distribution adaptation," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 2523911.

[24] Q. She, C. Zhang, F. Fang, Y. Ma, and Y. Zhang, "Multisource associate domain adaptation for cross-subject and cross-session EEG emotion recognition," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 2523911.

[25] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*. Berlin, Germany: Springer, 1992, pp. 162–190.

[26] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, no. 5, pp. 365–377, 2000.

[27] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2013, pp. 1247–1255.

[28] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, 2011.

[29] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10938–10947.

[30] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," 2020, *arXiv: 2009.04107*.

[31] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 14200–14213.

[32] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 4904–4916.

[33] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 9694–9705.

[34] C. Li, X. Lin, Y. Liu, R. Song, J. Cheng, and X. Chen, "EEG-based emotion recognition via efficient convolutional neural network and contrastive learning," *IEEE Sensors J.*, vol. 22, no. 20, pp. 19608–19619, Oct. 2022.

[35] M. Jiménez-Guarneros and G. Fuentes-Pineda, "CFDA-CSF: A multi-modal domain adaptation method for cross-subject emotion recognition," *IEEE Trans. Affective Comput.*, vol. 15, no. 3, pp. 1502–1513, Third Quarter 2024.

[36] C. Chen et al., "Comprehensive multisource learning network for cross-subject multimodal emotion recognition," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 9, no. 1, pp. 365–380, Feb. 2025.

[37] X. Gong, C. L. P. Chen, B. Hu, and T. Zhang, "CiABL: Completeness-induced adaptative broad learning for cross-subject emotion recognition with EEG and eye movement signals," *IEEE Trans. Affective Comput.*, vol. 15, no. 4, pp. 1970–1984, Fourth Quarter 2024.

[38] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
[39] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
[40] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Third Quarter 2014.
[41] H. Liu, Y. Zhang, Y. Li, and X. Kong, "Review on emotion recognition based on electroencephalography," *Front. Comput. Neurosci.*, vol. 15, 2021, Art. no. 84.
[42] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affective Comput.*, vol. 10, no. 3, pp. 417–429, Third Quarter 2019.
[43] J. Luo, Y. Tian, H. Yu, Y. Chen, and M. Wu, "Semi-supervised cross-subject emotion recognition based on stacked denoising autoencoder architecture using a fusion of multi-modal physiological signals," *Entropy*, vol. 24, no. 5, 2022, Art. no. 577.
[44] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=OOsR8BzCnl5
[45] Z. Han, F. Yang, J. Huang, C. Zhang, and J. Yao, "Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20675–20685.
[46] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. 7th Int. Conf. Artif. Neural Netw.*, Lausanne, Switzerland, Springer, 2005, pp. 583–588.

Qi Zhu (Member, IEEE) received the BS, MS degree, and PhD degrees from the Harbin Institute of Technology, in 2007, 2010, and 2014, respectively. He is a professor with the College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics. He has published more than 100 scientific articles in refereed international journal and conference proceedings. His interests include pattern recognition, feature extraction, and medical image analysis.

Ting Zhu received the BS degree in software engineering from the Nanjing University of Aeronautics and Astronautics, in Jul. 2021, where she is currently working toward the MS degree with the College of Artificial Intelligence. Her current research interests include machine learning and affective computing.

Lunke Fei (Senior Member, IEEE) received the PhD degree in computer science and technology from the Harbin Institute of Technology, Shenzhen, China, in 2016. He was a post-doctoral fellow with the Department of Computer and Information Science, University of Macau, Macau, China. He is currently an associate professor with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. His research interests include pattern recognition, biometrics, computer vision, and machine learning.

Chuhang Zheng received the BS degree in computer science and technology from Jimei University, in Jul. 2022. He is currently working toward the MS degree with the College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics. His current research interests include multi-modal learning and affective computing.

Wei Shao received the BS and MS degrees from the Nanjng University of Technology, Jiangsu, China, in 2009 and 2012, respectively, and the PhD degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2018. He is an associate professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His interests include machine learning and bioinformatics.

David Zhang (Life Fellow, IEEE) received the Graduate degree in computer science from Peking University, Beijing, China, the MSc degree in computer science and the first PhD degree from the Harbin Institute of Technology (HIT), Harbin, China, in 1982 and 1985, respectively, and the second PhD degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994. From 1986 to 1988, he was a post-doctoral fellow with Tsinghua University, Beijing, and then an associate professor with Academia Sinca, Beijing. He has been a chair professor with The Hong Kong Polytechnic University, where he has been the founding director with the Biometrics Research Centre (UGC/CRC) supported by the Hong Kong SAR Government, since 1998. He is currently the presidential chair professor with The Chinese University of Hong Kong (Shenzhen). He is also a visiting chair professor with Tsinghua University and an adjunct professor with Peking University, Shanghai Jiao Tong University, Shanghai, China, HIT, and the University of Waterloo. He has authored more than ten books and 200 journal articles. He is also a Croucher senior research fellow, a distinguished speaker of the IEEE Computer Society, and a fellow of International Association for Pattern Recognition (IAPR). He is an organizer of the International Conference on Biometrics Authentication. He was a book editor of the International Series on Biometrics (Springer) and an associate editor of more than ten international journals, including *IEEE Transactions and Pattern Recognition*. He is the founder and the editor-in-chief of the *International Journal of Image and Graphics*.

Daoqiang Zhang (Senior Member, IEEE) received the BS and PhD degrees in computer science from the Nanjing University of Aeronautics and Astronautics (NUAA), China, in 1999, and 2004, respectively. He joined the Department of Computer Science and Engineering of NUAA as a lecturer in 2004, and is a professor at present. His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has published more than 200 scientific articles in refereed international journals such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Medical Imaging*, *IEEE Transactions on Image Processing*, *Neuroimage*, *Human Brain Mapping*, *Medical Image Analysis*, and conference proceedings such as IJCAI, AAAI, NIPS, CVPR, MICCAI, KDD, with 12,000+ citations by Google Scholar. He was nominated for the National Excellent Doctoral Dissertation Award of China in 2006, won the best paper award or best student award of several international conferences such as PRICAI'06, STMI'12 and BICS'16, etc. He has served as a program committee member for several international conferences such as IJCAI, AAAI, NIPS, MICCAI, SDM, PRICAI, and ACML, etc. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence & Pattern Recognition Society of the China Computer Federation (CCF).