Multimodal Discriminative Network for Emotion Recognition Across Individuals

Minxu Liu, Donghai Guan, Chuhang Zheng, and Qi Zhu, *Member, IEEE*

***Abstract*—Multimodal emotion recognition is gaining significant attention for ability to fuse complementary information from diverse physiological and behavioral signals, which benefits the understanding of emotional disorders. However, challenges arise in multimodal fusion due to uncertainties inherent in different modalities, such as complex signal coupling and modality heterogeneity. Furthermore, the feature distribution drift in intersubject emotion recognition hinders the generalization ability of the method and significantly degrades performance on new individuals. To address the above issues, we propose a cross-subject multimodal emotion robust recognition framework that effectively extracts subject-independent intrinsic emotional identification information from heterogeneous multimodal emotion data. First, we develop a multichannel network with self-attention and cross-attention mechanisms to capture modality-specific and complementary features among different modalities, respectively. Second, we incorporate contrastive loss into the multichannel attention network to enhance feature extraction across different channels, thereby facilitating the disentanglement of emotion-specific information. Moreover, a self-expression learning-based network layer is devised to enhance feature discriminability and subject alignment. It aligns samples in a discriminative space using block diagonal matrices and maps multiple individuals to a shared subspace using a block off-diagonal matrix. Finally, attention is used to merge multichannel features, and multilayer perceptron is employed for classification. Experimental results on multimodal emotion datasets confirm that our proposed approach surpasses the current state-of-the-art in terms of emotion recognition accuracy, with particularly significant gains observed in the challenging cross-subject multimodal recognition scenarios.**

***Index Terms*—Attention mechanism, cross-subject, electroencephalogram (EEG), emotion recognition, multimodal fusion.**

I. Introduction

EMOTIONS play a crucial role in human society, constantly impacting individuals' psychological states and cognitive decision-making processes. Emotion state recognition technology [1], based on emotional signals such as electroencephalogram (EEG) [2], [3], can identify subtle emotional changes in individuals that may be key indicators of cognitive levels. In healthcare, by tracking changes in emotional responses over time, emotion recognition technology can also serve as an important tool for monitoring the progress of treatment for related brain disorders. However, because emotions represent real-time and intricate cognitive behaviors that exhibit variability among individuals and across different contexts, the development of an accurate and robust emotion recognition method continues to be a significant challenge.

Emotion recognition leverages two types of signals. Physiological signals encompass EEG, electromyography (EMG), and skin conductance response. Behavioral signals include facial expressions, speech patterns, and eye movements (EM) [4]. EEG is extensively utilized in emotion recognition due to its ease of acquisition and high temporal resolution [2], and various machine learning approaches focused on EEG have been developed [3], [5], [6], [7]. Song et al. [3] proposed a dynamic graph convolutional network (DGCNN) to extract the spatio-temporal representation from EEG signals. Du et al. [5] developed an attention-based multichannel LSTM (ATDD-LSTM) for emotion recognition. Zhang et al. [8] proposed an attention-based network, which synchronously obtained the global and local temporal features of EEG. However, the reliance on single-modal analysis in these methods may fall short of providing a comprehensive reflection of emotional states, which could result in potential limitations regarding accuracy and stability. Moreover, emotional responses are known to fluctuate considerably among individuals. Single-modal EEG emotion recognition methods often encounter difficulties in capturing this individual variability and in establishing a classification pattern that can be universally applied across diverse individuals.

Beyond single-modality analysis, the integration of multiple modalities usually provides a more comprehensive representation and complementary information for emotion states. Recently, an increasing number of studies combining EEG with other modalities such as EM and facial expressions have been proposed to improve the accuracy and robustness of emotion recognition [9], [10]. For example, synchronized alongside EEG and EM have proven to furnish complementary features beneficial for emotion analysis [11]. Liu et al. [12] developed a bi-modal deep auto-encoder (BDAE) [13] using EEG and EM modalities for emotion recognition. Wang et al. [14] employed transformer-based encoder with attention fusion to combine EEG and EM for improving the recognition

Received 7 July 2024; revised 19 October 2024 and 1 February 2025; accepted 11 March 2025. Date of publication 18 March 2025; date of current version 15 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62472220, Grant 62371234, and Grant 62076129, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20231438), and in part by Jiangsu Province 100 Foreign Experts Introduction Plan under Grant BX2022012. Recommended for acceptance by Associate Editor Jibin Wu. *(Corresponding authors: Donghai Guan; Qi Zhu.)*

The authors are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: dhguan@nuaa.edu.cn; zhuqi@nuaa.edu.cn).

Digital Object Identifier 10.1109/TCDS.2025.3552124

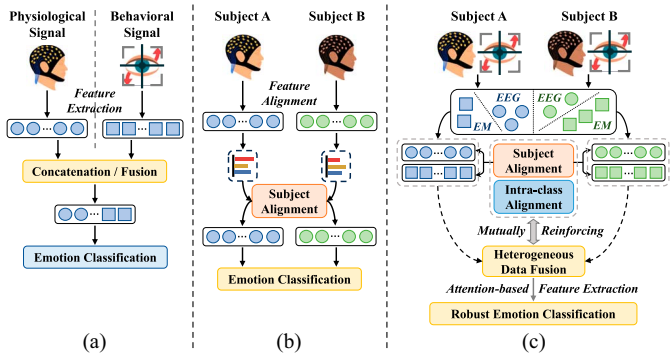

Fig. 1. Main difference between existing emotion recognition methods with the proposed method. (a) Multimodal emotion data fusion methods. (b) Cross-subject emotion recognition with single modal data. (c) Proposed multimodal discriminative network for cross-subject emotion recognition.

performance. Despite these advancements, the heterogeneity and redundancy of features remain significant influencing factors in multi-modal emotion recognition. Existing methods often struggle to effectively handle noise and coupled information within each modality, leading to variations in emotional expression and further complicating emotion computation.

In numerous real-world emotion recognition scenarios, there is frequently a discrepancy between the different individuals. The variability is attributed to the distinctive ways in which individuals express emotions, leading to the same emotion exhibiting a range of characteristics across different people. On widely used datasets such as SJTU Emotion EEG Dataset (SEED) [15], support vector machines (SVM) were used for both intrasubject and cross-subject experiments, and the effect showed a significant decline [16]. Recently, cross-subject studies based on EEG have received increasing attention [17], [18], [19], [20]. The two-phase prototypical contrastive domain generalization framework (PCDG) [18] utilized a contrastive learning method based on prototypical representations to narrow the representation differences between the subjects. The spatio-temporal self-constructing graph neural network (ST-SCGNN) [19] extracted activation and connectivity pattern features between different individuals and then combined them to leverage their complementary emotion-related information. Contrastive learning method for intersubject alignment (CLISA) [20] minimized intersubject differences by maximizing the similarity of EEG representations when subjects receive the same emotional stimuli versus different stimuli. The cross-subject setting is crucial for verifying the model's generalization capabilities, yet it also poses a challenge due to the varying data distributions among different individuals. Therefore, there is a pressing demand for the development of innovative approach to cross-subject multimodal emotion recognition, which not only strives to establish a unified representational space from disparate modalities and across individuals but also elevates the discriminative capacity of features to improve the accuracy of emotion recognition.

In this article, we propose a multichannel attention dual-cross network for multimodal cross-subject emotion recognition. Our approach effectively disentangles emotion-related information during the feature extraction phase and maps features across multiple individuals into a discriminative feature space, thereby mitigating the heterogeneity inherent in multimodal data across different individuals. Specifically, we start by employing self-attention and cross-attention to capture modality-specific and complementary features, respectively, from emotion signal. Simultaneously, a triplet contrastive loss is devised to optimize multichannel feature representations. Additionally, we design a network layer based on self-expression learning to enhance feature discriminability and subject-alignment. It utilizes class-based block diagonal matrices to align samples in a discriminative feature space and employs domain-based block off-diagonal matrices to map multiple individuals to a common individual space. Furthermore, we embed residual connections within the self-expression layer to prevent gradient vanishing or gradient exploding in the network. Finally, attention mechanisms are employed to fuse multi-channel features for extracting emotion-related features, which are then fed into the emotion classifiers and subject classifiers constructed using fully connected layers to achieve subject-independent emotion classification. The main differences between the existing emotion recognition strategies and our approach are shown in Fig. 1. Fig. 1(a) demonstrates the existing multimodal emotion fusion recognition strategy. Fig. 1(b) represents the existing cross-subject emotion recognition strategies. These existing methods cannot be directly applied to the multimodal cross-subject emotion recognition. Methodologically, our approach, shown in Fig. 1(c), is specifically designed for the challenging emotion recognition scenarios mentioned above, and it enhances the discriminative power of emotion features through self-expression module. It not only fuses the complementary information of heterogeneous modal data but also effectively calibrates the feature distribution across individuals, with these two processes mutually reinforcing each other. In brief, our method offers the following main advantages.

1) We propose a heterogeneous data fusion framework for multimodal cross-subject emotion recognition, capturing both modality-specific and complementary features with the attention machine, effectively disentanglement emotional information from physiological and behavioral signals.
2) A self-expression module is introduced to enhance feature discriminability and subject alignment using a deep learning network with block diagonal and block off-diagonal matrices.
3) We develop an attention mechanism based on contrastive learning to integrate key emotional features and use adversarial dual-discriminators to learn subject-invariance emotion features.
4) The experiment results on several EEG-based multimodal emotion datasets show that our proposed method surpasses state-of-the-art algorithms in emotion prediction, especially suitable for multimodal cross-subject scenarios.

The structure of the article is as follows: Section II reviews related works. Section III elaborates on the proposed method and its optimization. Section IV presents the experimental results.

Section V covers ablation studies and various visualizations. Finally, Section VI provides a summary.

II. Related Work

A. Multimodal Emotion Recognition

Recently, numerous multimodal emotion recognition methodologies have been explored. Ngai et al. [21] proposed a multibranch deep convolutional neural network, which fused EEG, EM, and facial expressions for emotion recognition. Huang et al.[22] applied canonical correlation analysis (CCA) [23] to fuse EEG and facial expression in a common feature space. Liu et al. [24] introduced a deep canonical correlation analysis (DCCA) incorporating an attention-based feature fusion method to enhance multi-modal emotion recognition. Jia et al. [9] developed a heterogeneous two-stream graph recurrent neural network for emotion classification to address the challenge of merging spatial-spectral-temporal domain features. Zhu et al. [25] proposed a confidence-aware fusion network of heterogeneous signals for emotion recognition, with a focus on systematically modeling uncertainty in fusion and capturing dynamic variations. Liu et al. [26] developed a multimodal mixed network for emotion recognition to use modal auxiliary tasks to achieve modal integration. Despite the progress in multimodal fusion techniques, challenges such as modal heterogeneity and pervasive noise within modalities continue to constrain the performance of emotion classification. In this article, we propose a robust emotion recognition network capable of simultaneously conducting modal representation alignment and multimodal fusion.

B. Unsupervised Domain Adaptation

The feature of source domain and target domain have different distributions for cross-subject emotion recognition, and it is necessary to utilize domain adaptation methods to alleviate gap to enhance recognition performance. For instance, transfer component analysis (TCA) [27] maps the features of different domains into a reproducing kernel Hilbert space (RKHS), reducing the distribution distance between the source domains and target domains. Gratton et al. [28] embed maximum mean discrepancy (MMD) [28] into neural networks as a metric for the distance between two domains, ensuring the similarity of distributions in a high-dimensional common projection space. Then, the adversarial-based domain adaptation mainly leverages the principles of generative adversarial networks (GANs), where the generator and discriminator adversarially generate sample features that possess both class discriminability and domain invariance simultaneously. The objective of model is learned with the minimization of the following loss function:

$$\min_{G_f,G_y,G_d} \mathcal{L}_{\mathrm{c}}(y, G_y(G_f(X))) - \beta \cdot \mathcal{L}_{\mathrm{d}}(d, G_d(G_f(X))) \quad (1)$$

where X is input data, $\mathcal{L}_{\mathrm{c}}$ is classifier, $\mathcal{L}_{\mathrm{d}}$ is domain classifier, G_f is feature generator, G_y and G_d are encoder of class-related feature and domain-related feature, β is regularization parameters, and y is class label and d is domain label. For example, Zhao et al. [29] developed multisource domain adversarial networks (MDAN), which optimized task-adaptive generalization bounds to deal with multisource domain adaptation. Pei et al. [30] proposed an uncertainty-induced transferability representation (UTR), which analyze channel transferability by source encoder uncertainty in unsupervised domain adaptive task. It is worth mentioning that these methods primarily focus on the calibration of distributions between different data domains, while neglecting the enhancement of feature discriminability. Motivated by adversarial-based domain adaptation [31], we develop a self-expression block constrain based to effectively alleviate the domain distribution difference across individuals while preserving emotion-related discriminability of classification features.

C. Multimodal and Cross-Subject Emotion Recognition

Many studies focus on multimodal emotion recognition and cross-subject emotion recognition. Recently, several methods have been proposed for learning subject-invariant emotional feature representations from multimodal data. However, the combination of these two tasks is not a simple overlay. For example, the distributional differences in cross-modality and cross-subject scenarios are multiplicatively amplified when combined. Gong et al. [32] proposed a multimodal emotion recognition model (CiABL), which synchronously aligned the features of two modalities in both source and target domains, and employs weighted representation distribution alignment to align both marginal and conditional probabilities simultaneously. Magdiel et al. [33] proposed the CFDA-CSF method for cross-subject emotion recognition using EEG and EM signals, which improved performance by aligning features both coarsely and finely and encouraging interclass separation by placing decision boundaries in low-density regions of the target domain. Chen et al. [34] introduced a comprehensive multisource learning network (CMSLNet), aligning multimodal features using a Huber loss-based consistency metric and fusing heterogeneous multimodal data through attention-based low-rank multimodal fusion in cross-subject emotion recognition. These approaches account for modality-specific differences across subjects when aligning heterogeneous multimodal data from multiple domains during the alignment phase. Our method, not only captures the relationships between different modalities during feature extraction but also enhances complementarity by emphasizing modality-specific characteristics. Through the self-expression module, our approach progressively integrates feature alignment into each layer of the feature extractor, achieving stepwise multimodal and cross-subject distribution alignment.

III. Proposed Method

In this section, we introduce our proposed multimodal cross-subject discriminative network, and the architecture of the proposed framework is shown in Fig. 2. The proposed method consists of four modules: multichannel feature embedding module, self-expression feature alignment module, contrastive-based modal fusion module, and classification module. The multichannel feature embedding module disentangles EEG and

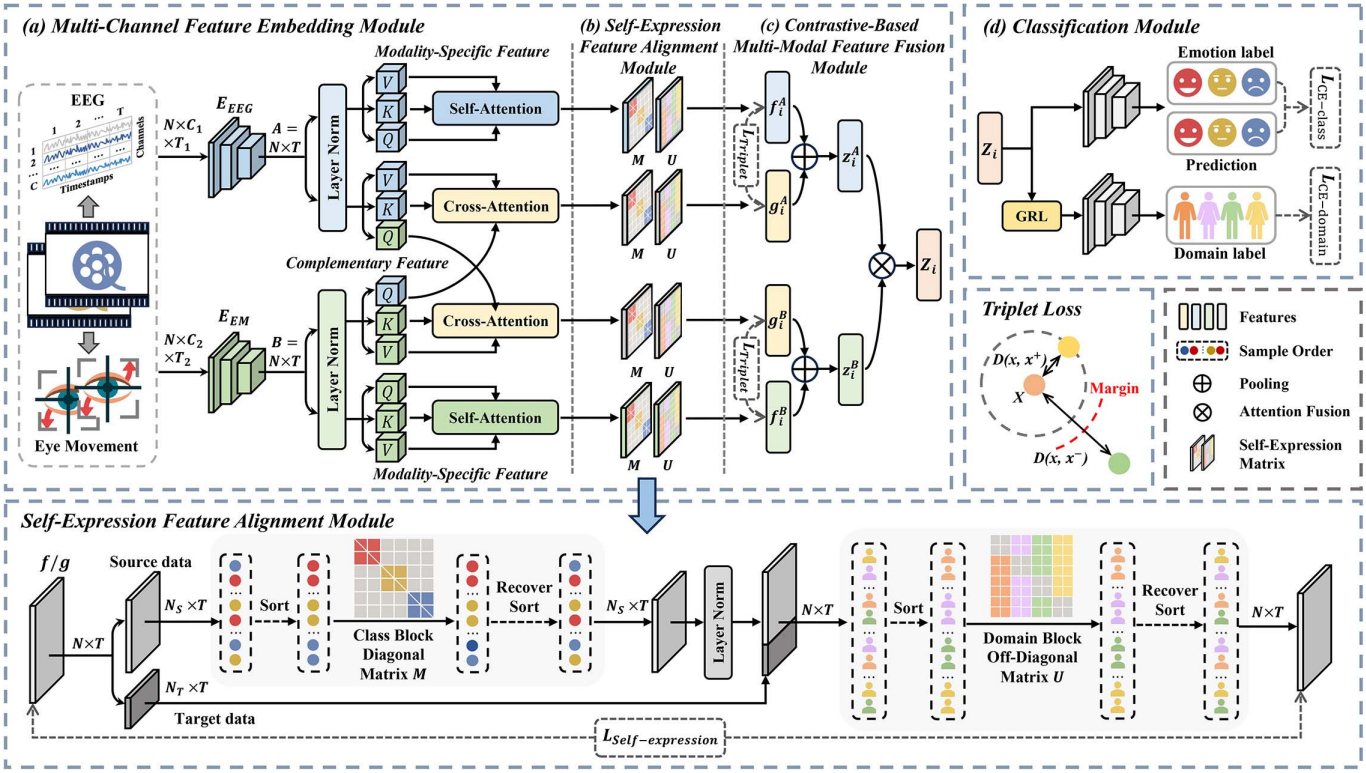

Fig. 2. Model architecture of the proposed multimodal cross-subject emotion recognition method comprises four modules: (a) multichannel feature embedding module, which includes (b) and (c) to extract emotion features from multimodal signals; (b) self-expression feature alignment module, embedded within the attention network, which enhances subject representation invariance and class discriminability; (c) contrastive-based modal fusion module, which fuses multichannel multimodal features; and (d) classification module, which uses a dual-discriminator structure to extract subject-invariant emotional features.

EM signals into modality-specific features and complementary features. The self-expression feature alignment module embeds a self-expression module in each layer of the attention network to align features into discriminative feature space. The contrastive learning-based multimodal feature fusion module further optimizes disentanglement using contrastive loss and integrates modal features from multiple channels using attention. Finally, emotion classification loss and subject discrimination loss are utilized to optimize features, further extracting subject-invariant emotion-related features.

A. Multichannel Feature Embedding Module

The model takes EEG and EM as input. We define the EEG as $D_{\text{EEG}} \in \mathbb{R}^{N\times C_1\times T_1}$ and the EM as $D_{\text{EM}} \in \mathbb{R}^{N\times C_2\times T_2}$, where N is sample number in a batch, $\{C_1, C_2\}$ is channels number of EEG and EM samples, and $\{T_1, T_2\}$ is sample length of EEG and EM. $N \in \{N_S, N_T\}$ is the sample number which includes source domains and target domain, denoting $D = [D_{\text{source}}, D_{\text{target}}]$. A linear layer is utilized to transform the dimensions of the two modalities into feature space with same dimensionality as $A \in \mathbb{R}^{N\times T}$ and $B \in \mathbb{R}^{N\times T}$.

We feed A and B into a multihead self-attention layer to extract modality-specific features. In the self-attention layer, queries (Q), keys (K), and values (V) for both modalities are computed using distinct linear projections

$$K_A = AW_A^K, Q_A = AW_A^Q, V_A = AW_A^V$$
$$K_B = BW_B^K, Q_B = BW_B^Q, V_B = BW_B^V \quad (2)$$

where W_K, W_Q, W_V are the parameter matrices utilized to generate keys, queries, and values. We determine the attention coefficients between brain regions by computing the scaled dot-product correlation between Q and K. These coefficients are then passed through the $\mathrm{SoftMax}$ function to obtain the attention weights. The final feature matrix with self-attention is produced by multiplying these attention weights by V

$$f_A^i = \text{selfAtt}(Q_A^i, K_A^i, V_A^i) = \text{softmax}\left(\frac{Q_A^i {K_A^i}^T}{\sqrt{d_1}}\right) V_A^i$$
$$f_B^i = \text{selfAtt}(Q_B^i, K_B^i, V_B^i) = \text{softmax}\left(\frac{Q_B^i {K_B^i}^T}{\sqrt{d_2}}\right) V_B^i \quad (3)$$

where $f_A^i \in \mathbb{R}^{N\times T}$ and $f_B^i \in \mathbb{R}^{N\times T}$ are the learned feature of EEG and EM after passing the self-attention layer, respectively, $\text{selfAtt}()$ denotes self-attention computing, and d is the normalization hyperparameter that $d_1 = d_2 = T$. We employ h heads for self-attention and each head can be denoted by $f^i = \text{selfAtt}(Q^i, K^i, V^i)$. Then, the output of the multihead self-attention is

$$f_X = (f_X^1 \| f_X^2 \| ... \| f_X^h)W + X \quad (4)$$

where h is the number of heads in self-attention layer, W is a linear transformation, $\|$ denotes the concatenation operation, and $X \in \{A, B\}$ is a set of EEG or EM features. The equation combines the outputs of multiple attention heads, and obtains representation applied a linear transformation and residual connection.

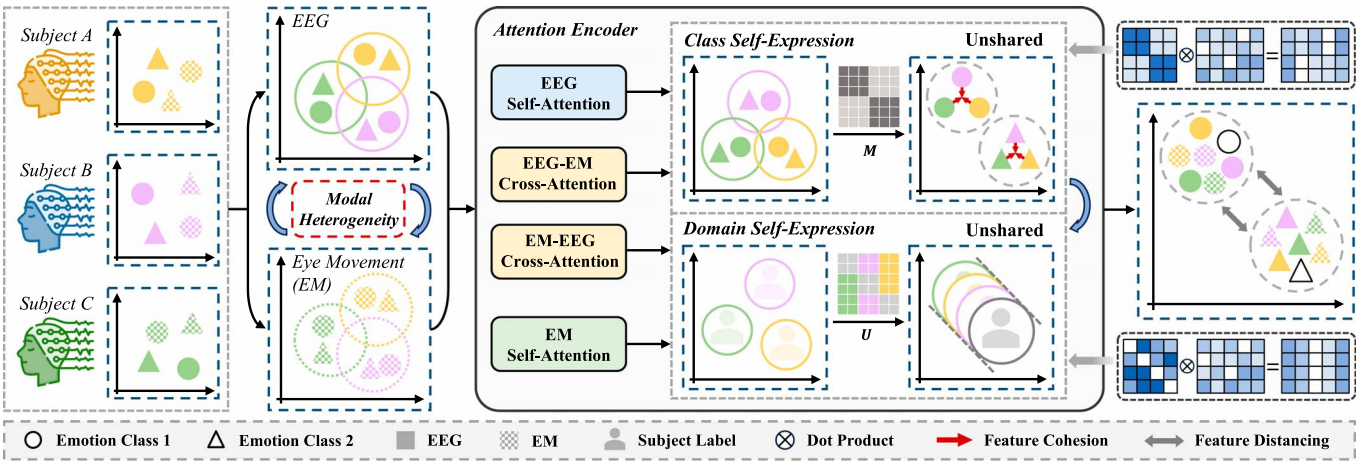

Fig. 3. Diagram of the proposed self-expression feature alignment module. Feature alignment is achieved by training two self-expression matrices: the block diagonal matrix M and the block off-diagonal matrix U. After passing through the attention encoder, the EEG and EM features of different subjects achieve enhanced class discrimination and subject invariance.

To disentangle the complementary features from the two modalities, we use cross-attention to perform cross-modal feature extraction. Compared with the calculation formula of self-attention in (3), cross-attention replaces the K and V of both EEG and EM with the other modality

$$g_A^i = \text{crossAtt}(Q_B^i, K_B^i, V_A^i) = \text{softmax}\left(\frac{Q_B^i {K_B^i}^T}{\sqrt{d_1}}\right) V_A^i$$
$$g_B^i = \text{crossAtt}(Q_A^i, K_A^i, V_B^i) = \text{softmax}\left(\frac{Q_A^i {K_A^i}^T}{\sqrt{d_2}}\right) V_B^i \tag{5}$$

where $g_A^i \in \mathbb{R}^{N\times T}$ and $g_B^i \in \mathbb{R}^{N\times T}$ are the new feature maps of EEG and EM after passing through the cross-attention layer, $\text{crossAtt}()$ denotes cross-attention computing, and the other parameters are consistent with the self-attention layer. The output of multihead cross-attention as follows:

$$g_X = (g_X^1 \| g_X^2 \| ... \| g_X^h)W + X. \tag{6}$$

B. Self-Expression Feature Alignment Module

Self-expression learning aims to represent a given sample using other samples, which optimizes features by constraining the difference between the features before and after the self-expression multiplication. We apply a set of learnable block diagonal matrices and a block off-diagonal matrix to represent the samples and propose two paradigms of self-expression.

1) *Class Self-Expression:* Achieved through a sparse block-diagonal matrix, it captures the category-discriminative representations within samples of the same emotion category.
2) *Domain Self-Expression:* Achieved through a block off-diagonal matrix, it models the relationships between features across different subjects, ensuring that the feature distribution of any subject can be represented by others, thus facilitating cross-subject alignment between the training and testing subjects.

To equip the model with both cross-modal alignment and cross-subject alignment, as depicted in Fig. 3, we propose a stepwise self-expression module. This layer can be embedded within each network to achieve two objectives simultaneously: 1) aligning same-category samples and eliminating noise of data; and 2) reducing differences across multiple domains in multimodal data.

The first self-expression module utilizes sparse block-diagonal matrices to promote the generation of compact, efficient representations corresponding to different categories, enhancing intraclass clustering. We introduce the self-expression coefficient matrix $M \in \mathbb{R}^{N_S \times N_S}$ as a diagonal block matrix, which facilitates preliminary sample alignment. After passing through this module, the process and constraints are as follows:

$$Y = \text{RecoverSort}(M \cdot \text{Sort}(X, l_{\text{cls}}))$$
$$X = Y, s.t. \text{diag}(M) = 0 \tag{7}$$

where $X \in \{f_A, g_A, f_B, g_B\}$ is the input of self-expression layer, l_{cls} is emotion category label of source subjects, $\text{Sort}(\cdot)$ is function to sort samples by emotion label, and $\text{RecoverSort}(\cdot)$ is function to restore the sample ordering before sorting.

The diagonal-block structure emerges in self-expression coefficients and each diagonal block corresponding to the corresponding category. To prevent trivial solutions and suppress mutual representation among different categories, we propose out-of-class mask matrix $Q \in \mathbb{R}^{N_S \times N_S}$ as

$$Q = 1_{N\times N} - \begin{pmatrix} 1_{n_1\times n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1_{n_c\times n_c} \end{pmatrix} + E_{N\times N} \tag{8}$$

where $n_i, i = 1 : c$ is the number of samples of class i. The self-expression and the diagonal-block constraint of M constitute the loss function of the self-expression layer

$$\mathcal{L}_{\text{ex-class}} = \|M\|_F^2 + \lambda_1 \sum_{i=1}^{n} \|MX - X\|_F^2 + \lambda_2 \|Q \odot M\|_F^2 \tag{9}$$

where λ_1, λ_2 are regularization parameters, $\|\cdot\|_F^2$ denotes the matrix Frobenius norm, and $\odot$ indicates the Hadamard product. In this equation, the first term prevents overfitting, the second term encourages M to obtain the self-expression ability, and the third term enhances the discriminative capability of M by minimizing the diagonal coefficients and out-of-class self-expression coefficients.

For the distribution difference between the target domain and the source domain, this article proposes utilizing self-expression module to introduce confusion among subjects. We assume that each subject can be constructed as a weighted combination of distributions from other subjects

$$X_k \longleftarrow \sum_{i=1}^{d} A_{ik} \cdot X_i \tag{10}$$

where X_k represents the kth subject, d is the number of subjects in X, and A_{ik} denotes the weight. Based on this assumption, we aim at using self-expression coefficient matrix U to closely approximate an block off-diagonal matrix, and the form of out-of-class mask matrix $K \in \mathbb{R}^{N\times N}$ as

$$K = \begin{pmatrix} 1_{n_1\times n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1_{n_d\times n_d} \end{pmatrix} \tag{11}$$

where n_i represents the number of samples in each subject. We combine the source and target domains as inputs of the self-expression layer, i.e., $X = (X_{\text{source}} \| X_{\text{target}})$, where X_{source} is Y in (7). We process self-expression operation as follows:

$$\begin{aligned} Y &= \text{RecoverSort}(U \cdot \text{Sort}(X, l_{\text{domain}})) \\ X &= Y, s.t. \text{diag}(U) = 0 \end{aligned} \tag{12}$$

where l_{domain} is the subjects label of X, and U is the second self-expression coefficient matrix that $U \in \mathbb{R}^{N\times N}$. The self-expression and its diagonal-block constraint constitute the loss

$$\mathcal{L}_{\text{ex-domain}} = \|U\|_F^2 + \lambda_3 \sum_{i=1}^{n} \|UX - X\|_F^2 + \lambda_4 \|K \odot U\|_F^2 \tag{13}$$

where λ_3, λ_4 are regularization parameters. Equations (9) and (13) represent intraclass alignment and subject distribution alignment, respectively, which complement each other and generate a discriminative feature space of multiple subjects. Each layer of the self-expression layer contains two consecutive self-expression processes. The former can eliminate differences between samples of the same type in multiple source domains. The latter can transfer the generalization from source domains to target domain, enabling the model to learn the feature distribution of the target domain. The loss function of the self-expression layer is formulated as follows:

$$\mathcal{L}_{\text{self-expression}} = \mathcal{L}_{\text{ex-class}} + \lambda_0 \mathcal{L}_{\text{ex-domain}} \tag{14}$$

where λ_0 is hyperparameter to adjust weights of $\mathcal{L}_{\text{ex-domain}}$. In the test stage, the model will skip the self-expression module.

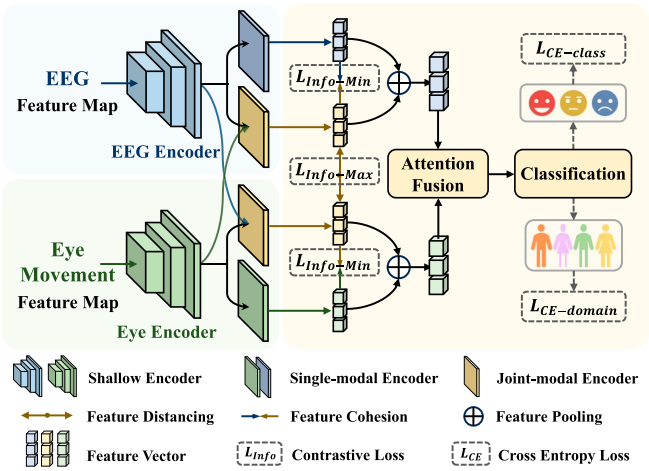

Fig. 4. Feature disentanglement network extracts high-dimensional features through the multichannel attention network with contrastive loss.

C. Contrastive-Based Multimodal Fusion

As shown in Fig. 4, through multimodal feature embedding and self-expression module, two modalities A and B obtain four features: f_A, g_A, f_B, and g_B, where f_A and f_B epresents the modality-specific features of EEG and EM, g_A and g_B represent the complementary features. For the aforementioned features, we utilize contrastive learning for further disentanglement. In each modality, since its mode-specific features and complementary features are derived from the same modality, we use the distance function to narrow their distance in the feature space. However, for the two sets of complementary features produced by the cross-attention shared by the two modalities, we need to increase their distance in the feature space. Since there are both positive and negative samples, we employ Triplet loss to constrain the feature contrast distance

$$\mathcal{L}_{\text{T}}(x, x^+, x^-) = \max\left(D(x^+, x) - D(x^-, x) + m, 0\right) \tag{15}$$

where $D(\cdot,\cdot)$ represents the Jensen-Shannon distance, x^+ is the feature that needs to be pulled closer to anchor x in their feature distance, x^- needs to be pulled farther from x, and m is boundary parameter. Therefore, we set the corresponding two sample pairs (x, x^+, x^-) for contrastive learning

$$\mathcal{L}_{\text{CL}} = \mathcal{L}_{\text{T}}(f_A, g_A, g_B) + \mathcal{L}_{\text{T}}(f_B, g_B, g_A). \tag{16}$$

Moreover, for the features f and g through contrastive learning, we employs attention mechanisms to fuse them. Attention mechanisms enable the model to automatically learn the weights of features, emphasizing information that is more relevant to current task during the fusion process. The final emotion-related features for classification by

$$Z = (\alpha_1 f_A + (1-\alpha_1) g_A) \| (\alpha_2 f_B + (1-\alpha_2) g_B) \tag{17}$$

where α_1 and α_2 denote the attention weights of EEG and EM signals, when combining the modality-specific and complementary features.

D. Classification Module

The model adopts an adversarial network architecture to further generate emotion-related features with subject invariance. Following the extraction of emotion feature, a domain discriminator and an emotion discriminator are integrated to obtain $\mathcal{L}_{\text{class}}$ and $\mathcal{L}_{\text{domain}}$. A gradient reversal layer (GRL) is utilized before the domain discriminator to ensure adversarial training, compelling shallow encoders to generate domain-invariant representations. The classification module is implemented by minimizing the following loss functions:

$$\mathcal{L}_{\text{CE}}(y,p) = -y\log p - (1-y)\log(1-p) \tag{18}$$

where y is the ground-truth label, and p is the predicted value. Consequently, $\mathcal{L}_{\text{class}} = \mathcal{L}_{\text{CE}}(y^{\text{class}}, C_{\text{class}}(z))$ and $\mathcal{L}_{\text{domain}} = \mathcal{L}_{\text{CE}}(y^{\text{domain}}, \text{GRL}(C_{\text{domain}}(z)))$ where C_{class} and C_{domain} are the classifiers for emotion labels and subject labels, respectively. The overall model is then optimized with the following objective function:

$$\mathcal{L} = \mathcal{L}_{\text{class}} + \mu\mathcal{L}_{\text{domain}} + \lambda\mathcal{L}_{\text{self-expression}} + \alpha\mathcal{L}_{CL} \tag{19}$$

where μ, λ, and α are hyperparameters to adjust the weights of $\mathcal{L}_{\text{domain}}$, $L_{\text{self-expression}}$ and $\mathcal{L}_{\text{CL}}$.

IV. Experiment

A. Data Materials and Preprocessing

We conducted the experiments on four emotion datasets to evaluate our proposed method: SEED [15], SEED-IV [4], SEED-V [35], and emotion analysis using physiological signals (DEAP) [36]. SEED, SEED-IV, and SEED-V are multimodal datasets with EEG and EM signals, while DEAP includes EEG and EOG signals. A summary description of dataset is shown in Table I.

SEED: This dataset is collected using the 62-channel ESI NeuroScan system and SMI eye-tracking glasses from twelve Chinese subjects who participated in an emotion elicitation experiment. Fifteen Chinese movie clips were used to induce positive, neutral, and negative three emotional categories, which lasted approximately 4 min for one experiment. A total of three sessions were recorded at different times.

SEED-IV: This dataset is collected from 15 subjects who watched 24 movie clips, each viewed three sessions at different times. One session of each subject contains 24 EEG trials, and the movie clips were selected to evoke happy, sad, fear, and neutral four emotional categories.

SEED-V: This dataset is collected from 16 subjects while they watched 15 video clips designed to induce happy, sad, neutral, fear, and disgusted five emotional categories. All subjects participated in three different sessions in three days, resulting in a total of 15 trials.

DEAP: This dataset is collected from 32 subjects while they watched 40 1-min video clips. All participants rated each video on a 1-9 scale with the indicators, i.e., arousal and valence.

We utilized the precomputed EEG and EM features available from the SEED, SEED-IV, and SEED-V datasets. Initially, the EEG raw data were downsampled from 1000 to 200 Hz. Next, differential entropy (DE) features were extracted using the short time Fourier transform (STFT) with a 4-s time window and no overlap, focusing on five specific frequency bands: δ (1–4 Hz), θ (4–8 Hz), α (8–14 Hz), β (14–31 Hz), and γ (31–50 Hz). The linear dynamic system (LDS) was used to eliminate the noise and artifacts for feature smoothing. Each EEG trial consists of 310 dimensions, derived from 62 channels across five frequency bands. Regarding EM, SMI BeGaze was used for feature extraction, which included 33 distinct EM features such as dispersion and other eye movement events. For all three datasets, we applied the standardization in EEG and EM features to extract initial features as the input of our method. For the DEAP dataset, we utilized 32-channel EEG and two-channel EOG signals for the experiment, and the dimensions of valence, and arousal were chosen as emotion assessment criteria. We downsample the EEG and EOG data from 512 to 128 Hz, and a 1 s nonoverlapping time window was implemented to generate samples. The power spectral density (PSD) features were extracted from precomputed EEG signals, and precomputed EOG is directly used as the eye-related features.

TABLE I
Summary of Datasets

Statistics	SEED		SEED-IV		SEED-V		DEAP	
	EEG	EM	EEG	EM	EEG	EM	EEG	EOG
Features/sample	310	33	310	31	310	33	128	256
Num. trails	540		1080		1080		1280	
Num. samples	30 312		37 425		29 168		76 800	
Num. subject	12		15		16		32	
Num. categories	3		4		5		2	

Note: Raw data points/trail means the average number of original data within a trail. Features/sample means the final extracted feature dimensions of a sample.

B. Experimental Settings and Comparison Methods

We experimentally validated the superiority of our proposed method in the cross-subject multimodal emotion recognition task. Beside, we conducted two additional tasks: intrasubject multimodal recognition and cross-subject unimodal recognition, to separately verify the effectiveness of our method in multimodal fusion and cross-subject alignment tasks.

1) Intrasubject Multimodal Setting: In this setting for SEED, SEED-IV, and SEED-V datasets, we divide the data of the first 60% trials into the training set and the remaining trials into the test set for each subject. For the DEAP dataset, 10-fold cross-validation is adopted. Specifically, we divide the samples of each subject into 10 subsets of equal size, using one subset as the test set and the other 9 as the training set in each fold. For our proposed method, we apply only the class self-expression paradigm, which aligns features at the category level to ensure consistent representation of samples within the same emotion category. The comparison includes MKL [37], CCA [23], KCCA [38], DCCA [39], DCCA-AM [24], BDAE [12], TMVC [40], ETF [14], MAET [41], MDDF [42], and CAFNet [25].

2) Cross-Subject Unimodal Setting: In this setting, we adopted the leave-one-subject-out experimental method in all

datasets. We utilize the single EEG modality of the source and target domains to perform our proposed method. For each subject in datasets, we use his/her EEG data from all trials as target samples, and the EEG data of the rest subjects as source samples. The average results of all subjects are recorded. The comparison includes TCA [27], KPCA [43], MMD [28], DAN [44], DANN [45], MADA [46], MDAN [29], DSAN [47], RSDA [48], FSA-TSP [49], DS-AGC [50], and BCJDA [51].

3) Cross-Subject Multimodal Setting: In this setting, a leave-one-subject-out cross-validation scheme is adopted to validate our overall model. For each subject of SEED, SEED-IV, and SEED-V datasets, we use the last 40% of his/her samples with EEG and EM modalities as target domain, and the first 60% of the multimodal data of the remaining subjects as source domain, which avoids the problem of repetitive stimuli in the training set and the test set on the basis of the reserve-one test method. For DEAP dataset, the leave-one-subject-out cross-validation strategy to validate our method. We selected some methods with relatively good performance in above two settings as the comparison methods. Additionally, CFDA-CSF [33], CiABL [32], MMDA [52], and CAFNet [25] are also included as the comparison methods in this experimental setting.

C. Implementation Details

Our proposed method is implemented in the public toolbox Pytorch, and all experiments were conducted with the NVIDIA GTX 3090 4-GPU. For the hyperparameters of the proposed model in all experiments, the number of attention layers of the multichannel feature extractor is set to 3, the dimensionality of the representation is set in the range of $\{128, 256, 512\}$, the dropout of self-attention and cross-attention feature extractor is set to 0.5, $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ is set to $\{0.8, 0.5, 1, 0.5, 1\}$, the boundary parameter m is set to 1.5, and the hyperparameters $\{\mu, \lambda, \alpha\}$ to adjust the weights of loss in (19) is set to $\{0.5, 0.8, 0.4\}$. During training, the model is trained using the Adaptive Moment Estimation (ADAM) optimizer with a batch size of $\{24, 32, 40, 32\}$ in SEED, SEED-IV, SEED-V, and DEAP datasets, respectively, the learning rate is set to $5e^{-3}$, and the grid search is adopted to find the optimal parameter combination for accuracy. Our code is available at https://github.com/xbrainnet/DCDNet.

D. Classification Performance on Emotion Recognition

The classification performance of our proposed method across three distinct scenarios, along with comparative results to other methods, is presented in Tables II–IV for the SEED, SEED-IV, SEED-V, and DEAP datasets, respectively, with values representing the mean/standard deviation for both accuracy, macro-F1, and F1 scores. As observed in Table II, our method surpasses all others in the intrasubject multimodal setting, with average accuracies of 91.79%, 89.94%, and 83.11% on the SEED, SEED-IV, and SEED-V datasets, respectively, while also achieving the highest Macro-F1 score. For the DEAP dataset, our proposed method achieves accuracies of 97.16% and 97.04% in the valence and arousal dimensions, respectively. These results highlight the effectiveness of our method in integrating multimodal heterogeneous emotional data by extracting both distinct and complementary features from various modalities. This can be attributed to the multichannel attention network's ability to disentangle modal-specific and complementary features from multimodal signals, enhancing modality fusion through contrastive learning. The block diagonal matrix in the self-expression module will not only enhance the similarity between samples of the same emotion category but also expand the discriminability of representation. Table III reveals that our method also excels in the cross-subject unimodal setting, due to the two-step alignment provided by the self-expression module, which enhances both subject alignment and discriminative feature representation. This demonstrates that our proposed self-expression module improves the class discriminability and subject invariance of the representations by enabling mutual representation among samples of the same class. Furthermore, the results presented in Table IV reveal that our method achieves the most distinguished recognition performance in cross-subject multimodal scenarios. This setting is particularly challenging and is a common occurrence in real-world applications, with our method leading in both accuracy and Macro-F1 scores. The efficacy of our method may be due to its concurrent handling of subject alignment and the amalgamation of heterogeneous data for the cross-subject multimodal emotion recognition task. Through the self-expression module embedded in the multichannel attention network, the multi-modal representation of emotion categories on different subjects is aligned. By contrastive learning, the interaction and fusion of multimodal representations are promoted, and the complementary connections are constructed while the modal heterogeneity is eliminated.

E. Feature Visualization

To verify the discriminability of the features, we use T-SNE to visualize the learned features for various methods, as illustrated in Fig. 5(f). Fig. 5(a) and 5(b) displays the T-SNE visualizations of the original EEG and EM features, respectively. From Fig. 5(c)–Fig. 5(e), the T-SNE visualizations are presented for the features after the fusion of the two modalities and subject calibration, respectively, showcasing the comparison methods and our approach. We visualize the features based on emotion labels and subject labels in the two rows of Fig. 5, respectively. It is evident that the original features shown in Fig. 5(a) and 5(b), whether EEG or EM signals, exhibit mixed distribution in categories but show clearer separations across individuals. The results indicate that our method can effectively differentiate between emotion categories. Using block diagonal matrices, we achieve cohesion of features within the same class, while the features exhibit similar distributions across different individuals. Specifically, all the three comparison methods employ the segregated strategy to multimodal fusion and subject alignment. While these approaches can enhance the separability of emotion categories, its classification performance is markedly inferior to that of our method. Our proposed approach demonstrates higher cohesion among samples within the same emotion class, more distinct boundaries between samples from

TABLE II
COMPARISONS OF DIFFERENT MULTIMODAL FUSION (MF) METHODS ON THE TASK OF INTRASUBJECT MULTIMODAL EMOTION RECOGNITION (%)

MF-Method	SEED		SEED-IV		SEED-V		DEAP(Valence)		DEAP(Arousal)	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	F1	Accuracy	F1
MKL	75.81/13.76	76.78/15.25	73.43/15.23	70.08/17.61	69.62/17.25	64.59/17.62	72.36/07.51	72.85/07.71	73.74/05.26	74.29/11.12
CCA	68.90/09.71	71.13/10.66	67.15/13.75	68.68/13.26	60.37/14.53	59.31/11.26	83.82/06.46	85.39/05.56	84.48/05.87	83.36/05.71
KCCA	76.75/09.46	75.39/10.42	71.52/12.72	67.71/13.59	66.61/13.16	62.62/13.09	84.10/04.64	84.78/05.66	84.12/05.41	85.63/06.34
DCCA	87.38/08.12	87.21/08.76	76.25/11.97	73.18/10.61	70.14/12.12	68.72/12.82	84.73/07.35	83.76/05.12	85.60/07.75	85.36/06.02
DCCA-AM	91.07/07.68	90.79/08.63	83.12/10.36	83.93/12.20	75.61/10.93	73.28/10.03	89.52/07.72	90.71/06.85	90.13/07.64	89.83/06.55
BDAE	88.24/08.93	88.05/09.07	84.86/12.02	82.68/11.68	80.07/06.61	77.49/07.12	91.93/06.53	91.42/05.25	92.32/06.09	92.11/05.93
TMVC	81.75/08.26	79.51/09.60	73.45/11.46	71.70/12.23	68.86/13.93	65.04/13.30	86.74/06.59	85.36/07.45	89.21/04.98	90.47/05.22
ETF	84.34/08.77	88.08/10.79	79.02/10.76	81.22/11.78	70.26/10.43	67.91/11.03	87.28/05.01	88.46/05.44	86.41/06.61	87.11/06.53
MAET	88.88/08.54	88.79/08.59	83.21/09.93	81.49/11.11	80.63/08.91	78.51/09.72	92.62/05.93	93.31/04.88	92.11/04.68	92.88/05.27
MDDF	90.13/09.69	90.92/09.21	82.49/10.97	80.68/09.23	81.02/09.85	80.80/10.90	93.15/04.51	94.14/04.21	93.62/04.12	94.23/04.58
CAFNet	91.33/07.86	90.97/08.13	89.16/07.44	88.03/08.49	82.35/09.66	**82.87/10.73**	95.25/03.84	94.50/03.97	94.89/03.07	95.55/03.30
Ours$_{MF}$	**91.79/07.02**	**91.04/07.86**	**89.94/06.76**	**88.84/08.80**	**83.11/09.26**	82.43/10.42	**97.16/02.16**	**98.35/01.86**	**97.04/02.31**	**96.98/01.88**

Note: The bold entries indicate the best result among all methods.

TABLE III
COMPARISONS OF DIFFERENT CROSS-SUBJECT (CS) METHODS ON THE TASK OF CROSS-SUBJECT UNIMODAL EMOTION RECOGNITION (%)

CS-Method	SEED		SEED-IV		SEED-V		DEAP(Valence)		DEAP(Arousal)	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	F1	Accuracy	F1
SO	64.09/12.24	70.52/11.73	63.26/14.53	60.28/10.61	52.19/14.74	50.38/15.89	51.36/05.12	52.61/07.81	52.64/06.85	53.18/08.02
TCA	73.47/11.46	74.78/12.31	66.43/11.63	67.09/11.89	58.63/14.35	53.21/13.23	52.51/08.26	53.75/11.81	54.09/09.14	54.98/10.76
KPCA	75.99/12.06	77.38/13.42	65.90/10.37	66.87/11.11	57.16/13.87	56.94/13.01	52.73/07.69	53.13/09.22	53.34/08.48	54.36/08.40
MMD	84.96/10.96	83.97/09.89	77.49/08.20	78.04/08.80	75.76/09.70	71.31/10.12	56.22/08.64	57.63/10.59	58.82/13.34	59.08/12.16
DAN	82.39/09.07	81.18/09.85	74.24/10.05	75.82/09.32	76.39/11.32	74.27/12.37	54.87/07.51	56.70/08.81	56.31/09.56	57.34/10.03
DANN	86.23/07.97	85.59/07.85	81.24/07.44	83.63/08.48	77.93/11.48	75.38/11.16	59.26/06.33	59.15/07.47	61.43/10.18	63.35/11.64
MADA	83.32/08.59	84.00/09.07	72.34/09.11	73.51/08.78	70.57/10.07	69.21/11.43	57.42/07.29	56.70/08.12	58.93/11.75	58.76/13.28
MDAN	82.47/10.38	83.31/10.36	72.48/07.78	74.10/08.20	69.50/10.71	66.69/10.09	57.33/08.03	58.23/09.68	57.98/10.36	59.35/12.73
DSAN	81.13/09.69	80.72/09.21	76.49/10.97	75.69/09.23	78.01/11.34	76.78/12.65	60.24/07.67	61.76/09.64	61.38/09.89	62.14/11.41
RSDA	83.52/08.70	82.50/09.64	82.49/11.13	80.68/09.83	77.24/10.72	77.03/10.15	61.53/07.93	62.15/08.47	62.43/10.18	63.35/11.64
FSA-TSP	91.79/07.50	91.76/07.54	85.45/09.81	**85.33/09.22**	-	-	-	-	-	-
DS-AGC	87.37/06.19	-	66.00/07.93	-	59.40/09.99	-	-	-	-	-
BCJDA	90.25/07.27	90.89/08.02	81.40/08.51	81.72/07.96	78.05/11.45	77.64/11.59	62.98/06.38	62.10/08.92	64.32/09.15	64.19/11.42
Ours$_{CS}$	**92.52/07.13**	**93.10/07.47**	**85.63/07.54**	84.87/07.32	**80.63/09.43**	**79.14/10.17**	**65.16/05.38**	**64.75/09.43**	**66.31/08.60**	**65.72/10.42**

Note: The bold entries indicate the best result among all methods.

TABLE IV
COMPARISON OF DIFFERENT COMBINATORIAL METHODS ON THE TASK OF CROSS-SUBJECT MULTIMODAL EMOTION RECOGNITION (%)

MF-Method	CS-Method	SEED		SEED-IV		SEED-V		DEAP(Valence)		DEAP(Arousal)	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	F1	Accuracy	F1
KCCA	TCA	74.52/10.62	75.62/10.36	61.52/09.96	62.72/10.57	66.13/17.96	64.35/16.79	56.13/09.57	61.19/11.82	57.03/07.46	59.46/13.05
	MMD	72.33/12.63	73.08/18.85	68.53/09.05	66.87/08.29	73.66/15.12	71.09/16.41	62.13/08.34	64.19/10.71	63.03/09.77	64.46/14.28
	MDAN	77.52/10.52	78.52/10.63	76.69/07.26	76.71/07.71	71.50/14.71	70.69/15.09	61.84/08.08	63.78/11.69	61.78/10.19	64.51/12.70
	DANN	76.82/11.66	77.10/11.05	75.92/08.16	76.94/08.10	74.46/15.75	74.66/14.53	63.86/07.57	65.05/11.49	63.19/09.89	65.58/11.87
DCCA	TCA	78.47/11.81	80.01/10.55	75.32/15.90	74.12/15.49	58.72/10.61	56.64/10.51	59.73/12.65	63.08/10.65	62.21/08.88	61.05/13.63
	MMD	72.33/12.63	73.08/18.85	68.53/09.05	66.87/08.29	73.66/15.12	71.09/16.41	65.70/07.41	67.37/10.30	64.47/09.71	66.98/12.07
	MDAN	77.52/10.52	78.52/10.63	76.69/07.26	76.71/07.71	71.50/14.71	70.69/15.09	64.18/08.53	65.39/11.56	63.29/10.40	64.65/13.65
	DANN	76.82/11.66	77.10/11.05	75.92/08.16	76.94/08.10	74.46/15.75	74.66/14.53	66.29/09.17	68.64/09.64	67.11/09.38	66.72/12.34
MDDF	TCA	80.26/10.89	81.23/11.08	67.45/13.01	67.02/14.85	70.38/11.35	71.32/11.08	59.50/11.41	63.65/10.06	64.83/09.16	67.45/12.59
	MMD	90.77/06.03	89.83/07.13	83.72/08.94	85.06/08.87	77.23/12.50	78.35/14.24	68.53/06.94	69.19/09.65	67.73/10.37	68.62/11.51
	MDAN	90.89/06.56	91.92/07.77	79.10/06.84	80.24/07.35	78.74/14.05	77.38/13.57	67.74/09.45	68.60/10.89	65.56/11.75	67.10/13.52
	DANN	91.50/07.29	91.32/08.05	84.02/07.80	83.98/06.92	82.53/13.35	83.23/12.50	68.15/08.77	68.97/11.08	68.85/09.72	69.75/11.44
RSDA	TCA	75.14/10.92	76.33/10.02	62.90/08.81	64.07/09.63	69.23/14.95	70.09/15.78	58.51/11.19	62.56/12.27	62.05/09.57	62.42/12.50
	MMD	81.08/10.63	81.79/11.83	78.89/08.72	78.09/09.24	75.34/13.28	76.53/14.35	67.91/07.21	68.61/11.70	68.41/11.76	67.32/11.98
	MDAN	83.43/09.81	82.23/10.12	79.72/07.72	79.34/09.05	76.27/14.27	77.23/15.07	67.24/08.33	67.90/11.02	66.75/12.50	66.92/13.19
	DANN	82.90/09.62	83.82/10.61	80.25/06.82	81.04/07.82	81.78/12.10	82.23/13.64	68.02/09.34	69.59/09.63	69.31/10.56	68.53/12.16
CFDA-CSF		90.47/05.08	90.37/05.18	80.63/06.14	80.38/06.16	84.89/11.30	84.42/11.55	-	-	-	-
CiABL		64.79/12.86	75.03/10.88	61.00/12.84	74.02/08.72	73.19/07.55	83.40/04.70	-	-	-	-
MMDA		94.82/02.41	94.75/02.41	85.54/08.11	85.28/08.05	86.77/10.32	86.56/10.42	-	-	-	-
CAFNet		92.62/07.03	92.21/07.82	86.14/07.34	85.25/06.50	85.76/11.21	86.65/10.66	72.22/06.68	71.81/08.08	70.69/08.94	**76.83/10.31**
Ours		**94.93/05.75**	**94.95/06.96**	**89.31/06.53**	**90.96/05.70**	**88.76/09.16**	**90.09/07.51**	**73.51/06.32**	**72.87/08.17**	**73.63/08.18**	75.05/09.62

Note: The bold entries indicate the best result among all methods.

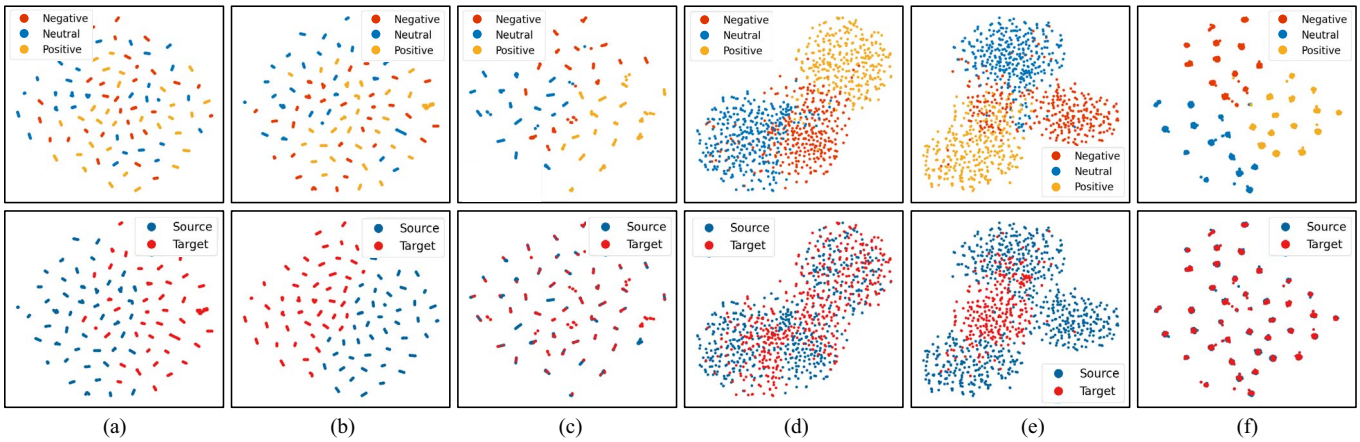

Fig. 5. T-SNE visualization for the features learned by different methods on SEED dataset under the multimodal cross-subject experimental setting. The first row shows the representation distribution with emotion labels, while the second row shows representation distribution with subjects (source and target) labels. (a) Original EEG. (b) Original EM. (c) DCCA + MMD. (d) DCCA + DANN. (e) MDDF + DANN. (f) Ours.

TABLE V
ABLATION STUDY FOR CROSS-SUBJECT MULTIMODAL CLASSIFICATION PERFORMANCE (MEAN/STD) (%)

Method			SEED		SEED-IV		SEED-V		DEAP(Valence)		DEAP(Arousal)	
CFD	SE	$\mathcal{L}_{CL}$	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	F1	Accuracy	F1
×	×	×	83.50/09.01	81.03/08.56	71.24/10.12	70.78/09.65	70.89/09.85	70.12/09.47	63.41/07.83	62.15/08.92	64.33/07.45	63.50/09.33
√	×	×	90.33/07.50	89.80/07.97	85.74/07.24	84.93/07.87	83.65/09.14	82.89/08.92	69.73/08.04	68.89/07.55	68.55/08.63	69.20/08.79
×	√	×	86.72/07.33	87.87/08.19	80.92/09.02	80.15/09.13	78.46/09.86	77.78/08.95	66.22/06.94	65.01/08.11	65.03/07.66	64.89/08.23
×	×	√	85.40/08.97	85.22/08.50	78.85/08.91	78.10/09.25	76.34/09.78	75.89/09.18	64.87/07.02	64.13/08.43	63.72/08.15	63.59/08.86
√	√	×	91.44/06.75	92.67/07.21	88.03/06.59	88.55/07.04	86.72/08.94	87.03/08.85	72.15/06.63	71.89/07.54	71.45/07.91	71.62/08.04
√	×	√	90.76/06.81	91.53/08.05	86.91/07.01	87.23/07.92	85.12/08.96	84.78/08.11	70.81/07.23	70.44/08.02	70.23/07.73	70.65/08.37
×	√	√	89.61/07.93	89.14/08.36	84.72/07.68	84.11/08.19	82.56/09.32	81.89/09.13	69.04/06.98	68.75/08.33	68.45/08.12	68.89/08.64
W/o self-attention			92.89/06.35	93.07/07.27	88.87/06.23	89.12/07.18	87.61/09.14	87.85/08.64	72.76/06.85	72.32/08.15	71.85/08.02	72.04/08.27
W/o cross-attention			92.11/07.47	92.65/07.63	88.12/06.91	88.34/07.23	86.78/09.42	86.65/08.73	71.89/07.23	71.50/08.02	71.25/07.94	71.40/08.18
JS $\Longrightarrow$ Euclid			91.74/08.04	92.02/07.94	87.59/06.74	87.80/07.85	85.92/09.13	85.56/08.59	71.12/07.45	70.75/08.11	70.53/07.73	70.79/08.45
JS $\Longrightarrow$ Consine			93.39/07.21	93.68/07.81	89.03/06.28	89.31/07.14	88.02/08.96	87.72/08.41	73.02/06.67	71.88/08.53	72.54/07.96	72.79/08.22
Ours			**94.22/05.75**	**94.95/06.96**	**89.31/06.53**	**90.96/05.70**	**88.76/09.16**	**90.09/07.51**	**73.51/06.32**	**72.87/08.17**	**73.63/08.18**	**75.05/09.62**

Note: The bold entries indicate the best result among all methods.

different classes, and a significant degree of subject representation invariance. This is due to our method simultaneously addressing the fusion of heterogeneous emotional features and the calibration of features across individuals, with these two aspects mutually reinforcing each other to achieve superior performance.

V. DISCUSSION

A. Ablation Study

The main method proposed in the article comprises three parts: the cross-attention feature disentanglement (CFD) module, the self-expression (SE) module, and the contrastive loss ($\mathcal{L}_{CL}$) module. To rigorously evaluate their contributions, we conducted ablation studies by systematically removing each module. As shown in Table V, removing any individual module leads to a significant performance drop, with accuracy reductions of up to 10.72% (SEED), 18.07% (SEED-IV), 17.87% (SEED-V), 10.10% (DEAP Valence), and 9.30% (DEAP Arousal), highlighting the necessity of each component. It is evident from the table that utilizing paired combinations of modules, as shown in Table V, yields superior classification performance compared with using individual modules alone. This finding serves to underscore the significance of each module within our proposed method.

To further assess the role of self-attention and cross-attention, and to test the adaptability of the self-expression module, we replaced the attention mechanisms with equivalent multilayer perceptrons (MLPs). For self-attention, we used a direct MLP replacement, while for cross-attention, we used two MLPs to extract features from different modalities, concatenated the results, and mapped them to the same dimensionality as the original cross-attention output. The results in Table V indicate that self-attention enhances representation learning, improving accuracy by 0.44%–1.78%, while cross-attention is crucial for multimodal fusion, boosting accuracy by 1.19%–2.38%. Additionally, we replaced the Jensen-Shannon distance in the contrastive learning module with Euclidean and cosine distances. The results show that Jensen-Shannon distance performs best, likely because Euclidean distance struggles with large feature value variations, while cosine distance is more suited for high-dimensional sparse data or normalized vectors.

Moreover, as shown in Table VI, varying the number of attention layers confirms that three layers strike the best balance between complexity and generalization. Reducing to one or two layers results in performance drops of 0.89%–1.34%, whereas

TABLE VI
ABLATION STUDY FOR THE NUMBER OF ATTENTION LAYERS OF THE MULTICHANNEL FEATURE EXTRACTOR (MEAN/STD) (%)

Method	SEED		SEED-IV		SEED-V		DEAP(Valence)		DEAP(Arousal)	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	F1	Accuracy	F1
$Layer_{att} = 1$	93.05/07.12	93.25/07.89	88.32/06.85	88.45/07.34	86.96/08.97	87.18/08.52	72.35/07.02	72.05/08.22	71.53/08.14	71.77/08.39
$Layer_{att} = 2$	93.22/06.97	93.40/07.68	88.59/06.53	88.75/07.11	87.21/08.74	87.15/08.33	72.45/06.81	72.16/08.10	71.71/07.98	71.95/08.31
$Layer_{att} = 4$	91.89/08.52	92.12/08.43	86.92/07.23	86.54/07.98	85.12/09.62	84.20/09.01	69.76/87.58	69.92/08.45	69.53/08.27	69.85/08.54
$Layer_{att} = 3$	**94.22/05.75**	**94.95/06.96**	**89.31/06.53**	**90.96/05.70**	**88.76/09.16**	**90.09/07.51**	**73.51/06.32**	**72.87/08.17**	**73.63/08.18**	**75.05/09.62**

Note: The bold entries indicate the best result among all methods.

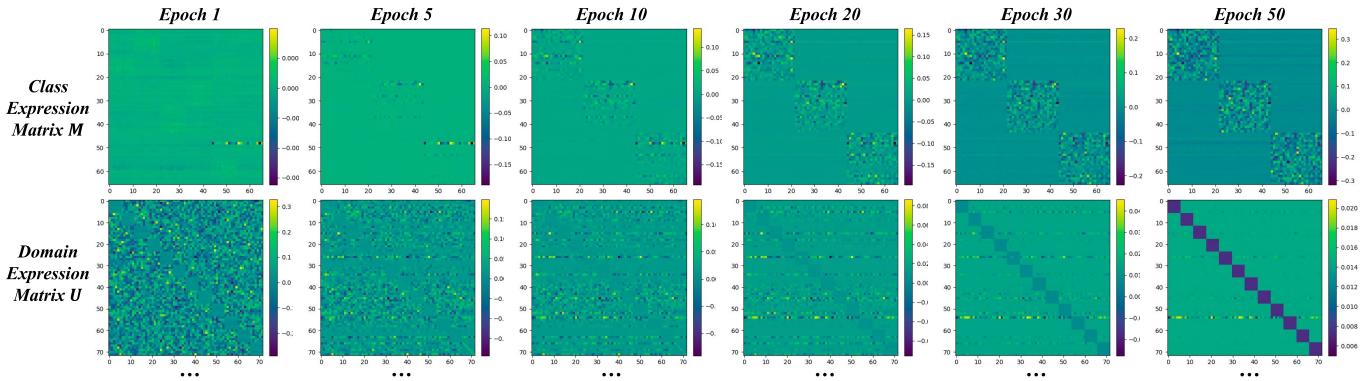

Fig. 6. Visualization for class self-expression matrixes and domain self-expression matrixes learned by the proposed method.

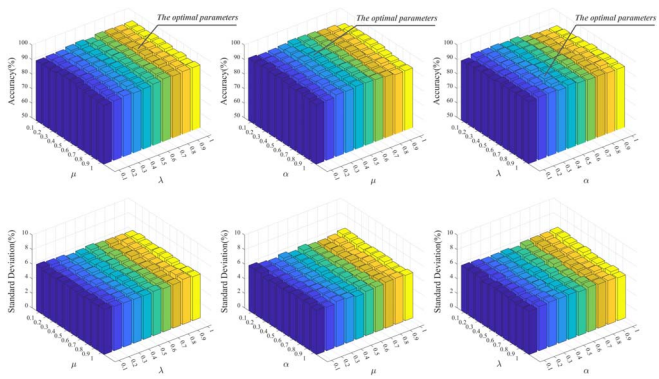

Fig. 7. Performance of classification accuracy of the proposed method under different parameters.

increasing to four layers introduces overfitting, causing slight declines across datasets. These findings validate the effectiveness of our design choices.

B. Self-Expression Functional Analysis

To validate the role of the self-expression module, we randomly extracted a layer from the self-expression layer of the trained model and visualized two self-expression matrices. As shown in Fig. 6, with an increase in training epochs, the self-expression matrices gradually evolve from initial chaotic forms to clarity. The class self-expression matrix within the model gradually learns representations within different classes, forming a clear diagonal block matrix, indicating its ability to impart discriminability on features regarding classes. Meanwhile, the domain self-expression matrix gradually learns commonalities between different subjects, forming a block off-diagonal matrix. Furthermore, the self-expression matrices from different layers exhibit slight differences, indicating variances in adapting to noise distributions across layers. These results underscore the effectiveness of the self-expression layer in this model, enabling feature learning with both class discriminability and domain invariance.

C. Parameters Analysis

The objective function in (19) introduces three regularization parameters μ, λ, and α in our proposed method. These regularization parameters are utilized to balance the contributions of domain classification constraint, self-expression constraint, and contrastive constraint. To explore the classification accuracy under different parameter settings, we conducted experiments on the SEED dataset using a cross-subject multimodal experimental setting. The regularization parameters μ, λ, α are each selected from the set $\{0.1, 0.2, ..., 1\}$. We adopt grid search to determine the optimal parameters through the test set accuracy across various hyperparameter combinations of these parameters, and the optimal values are $\mu = 0.5$, $\lambda = 0.8$, $\alpha = 0.4$. For visualization purpose, we held one parameter constant at its optimal value and systematically varied the other two parameters in pairs to observe the resulting changes in accuracy. The results presented in Fig. 7 show that our proposed method is not sensitive to these parameters.

VI. CONCLUSION

In this article, we propose a multichannel framework for heterogeneous feature integration and distribution calibration

for cross-subject multimodal emotion recognition. Our method embeds the fusion of heterogeneous emotional data and subject calibration within a unified recognition framework, where these processes mutually enhance each other to elevate the performance of the recognition task. Specifically, our proposed method first leverages self-attention and cross-attention mechanisms to respectively capture specific and complementary features across modalities. Second, we incorporate a contrastive loss function into the multichannel attention network, which significantly enhances feature extraction across different channels and achieves disentanglement of emotion information. Additionally, a self-expression learning-based network layer is proposed to enhance feature discriminability and subject-agnosticism. It utilizes block diagonal matrices to align samples in a discriminative space and block off-diagonal matrices to map multiple individuals to common subspace. Finally, we employ an adversarial dual-discriminator mode to constrain the model to generate subject-agnostic discriminative features. Experimental results from a range of multimodal emotion recognition datasets indicate that our proposed method consistently achieves superior accuracy compared with current methodologies. Significantly, our model's performance is particularly robust while recognizing emotions in cross-subject multimodal scenario.

REFERENCES

[1] D. O. Bos et al., "EEG-based emotion recognition," *Influence Vis. Auditory Stimuli*, vol. 56, no. 3, pp. 1–17, 2006.

[2] S. Slobounov, M. Hallett, S. Stanhope, and H. Shibasaki, "Role of cerebral cortex in human postural control: an EEG study," *Clin. Neurophysiol.*, vol. 116, no. 2, pp. 315–323, 2005.

[3] A. V. Phan, M. Le Nguyen, Y. L. H. Nguyen, and L. T. Bui, "DGCNN: A convolutional neural network over large-scale labeled graphs," *Neural Netw.*, vol. 108, pp. 533–543, Sep. 2018.

[4] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2018.

[5] X. Du et al., "An efficient LSTM network for emotion recognition from multichannel eeg signals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1528–1540, Jul./Sep. 2022.

[6] Q. Li, T. Zhang, C. P. Chen, K. Yi, and L. Chen, "Residual GCB-Net: Residual graph convolutional broad network on emotion recognition," *IEEE Trans. Cogn. Developmental Syst.*, vol. 15, no. 4, pp. 1673–1685, Dec. 2023.

[7] Y. Li et al., "A novel bi-hemispheric discrepancy model for EEG emotion recognition," *IEEE Trans. Cogn. Developmental Syst.*, vol. 13, no. 2, pp. 354–367, Jun. 2021.

[8] Y. Zhang, H. Liu, D. Zhang, X. Chen, T. Qin, and Q. Zheng, "EEG-based emotion recognition with emotion localization via hierarchical self-attention," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2458–2469, Jul./Sep. 2023.

[9] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, "HetEmotionNet: two-stream heterogeneous graph recurrent neural network for multimodal emotion recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1047–1056.

[10] Y. Yang, Q. Gao, Y. Song, X. Song, Z. Mao, and J. Liu, "Investigating of deaf emotion cognition pattern by EEG and facial expression combination," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 2, pp. 589–599, Feb. 2022.

[11] X. Wu, W.-L. Zheng, Z. Li, and B.-L. Lu, "Investigating eeg-based functional connectivity patterns for multimodal emotion recognition," *J. Neural Eng.*, vol. 19, no. 1, 2022, Art. no. 016012.

[12] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Neural Inf. Process.: 23rd Int. Conf. (ICONIP)*, Kyoto, Japan: Springer, Oct. 2016, pp. 521–529.

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

[14] Y. Wang, W.-B. Jiang, R. Li, and B.-L. Lu, "Emotion transformer fusion: complementary representation properties of EEG and eye movements on recognizing anger and surprise," in *Proc. IEEE Int. Conf. Bioinf. Biomedicine (BIBM)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 1575–1578.

[15] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[16] L.-M. Zhao, X. Yan, and B.-L. Lu, "Plug-and-play domain adaptation for cross-subject eeg-based emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 1, 2021, pp. 863–870.

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

[18] H. Cai and J. Pan, "Two-phase prototypical contrastive domain generalization for cross-subject eeg-based emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 1–5.

[19] J. Pan et al., "ST-SCGNN: a spatio-temporal self-constructing graph neural network for cross-subject eeg-based emotion recognition and consciousness detection," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 2, pp. 777–788, Feb. 2024.

[20] 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. Affect. Comput.*, vol. 14, no. 3, pp. 2496–2511, Jul./Sep. 2023.

[21] W. K. Ngai, H. Xie, D. Zou, and K.-L. Chou, "Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources," *Inf. Fusion*, vol. 77, pp. 107–117, 2022.

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

[23] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics: Methodology and Distribution*, New York, NY, USA: Springer, 1992, pp. 162–190.

[24] 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.

[25] Q. Zhu, C. Zheng, Z. Zhang, W. Shao, and D. Zhang, "Dynamic confidence-aware multi-modal emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1358–1370, Jul./Sep. 2024.

[26] F. Liu, P. Yang, Y. Shu, F. Yan, G. Zhang, and Y.-J. Liu, "Emotion dictionary learning with modality attentions for mixed emotion exploration," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1289–1302, Jul./Sep. 2024.

[27] 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.

[28] 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. 1, pp. 723–773, 2012.

[29] 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.*, vol. 31, 2018.

[30] J. Pei, Z. Jiang, A. Men, L. Chen, Y. Liu, and Q. Chen, "Uncertainty-induced transferability representation for source-free unsupervised domain adaptation," *IEEE Trans. Image Process.*, vol. 32, pp. 2033–2048, 2023.

[31] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2015, pp. 1180–1189.

[32] X. Gong, C. 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. Affect. Comput.*, vol. 15, no. 4, pp. 1970–1984, Oct./Dec. 2024.

[33] M. Jiménez-Guarneros and G. Fuentes-Pineda, "CFDA-CSF: A multimodal domain adaptation method for cross-subject emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 4, pp. 1970–1984, Oct./Dec. 2024.

[34] C. Chen et al., "Comprehensive multisource learning network for cross-subject multimodal emotion recognition," *IEEE Trans. Emerg. Topics in Comput. Intell.*, vol. 9, no. 1, pp. 365–380, Feb. 2025.
[35] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 2, pp. 715–729, Jun. 2022.
[36] S. Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.
[37] Q. Cai, G.-C. Cui, and H.-X. Wang, "EEG-based emotion recognition using multiple kernel learning," *Mach. Intell. Res.*, vol. 19, no. 5, pp. 472–484, 2022.
[38] K. Fukumizu, F. R. Bach, and A. Gretton, "Statistical consistency of kernel canonical correlation analysis," *J. Mach. Learn. Res.*, vol. 8, no. 2, 2007, pp. 361–383.
[39] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2013, pp. 1247–1255.
[40] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551–2566, Feb. 2023.
[41] W.-B. Jiang, X.-H. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 5975–5984.
[42] 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. 20707–20717.
[43] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artif. Neural Netw.*, Berlin, Heidelberg, Germany: Springer, 1997, pp. 583–588.
[44] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2015, pp. 97–105.
[45] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.
[46] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
[47] Y. Zhu et al., "Deep subdomain adaptation network for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1713–1722, Apr. 2021.
[48] X. Gu, J. Sun, and Z. Xu, "Unsupervised and semi-supervised robust spherical space domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1757–1774, Mar. 2024.
[49] M. Jiménez-Guarneros and G. Fuentes-Pineda, "Cross-subject EEG-based emotion recognition via semi-supervised multi-source joint distribution adaptation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023, Art. no. 2523911.
[50] W. Ye et al., "Semi-supervised dual-stream self-attentive adversarial graph contrastive learning for cross-subject eeg-based emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 16, no. 1, pp. 290–305, Jan./Mar. 2025.
[51] Y. Shao, Y. Zhou, P. Gong, Q. Sun, and D. Zhang, "A dual-adversarial model for cross-time and cross-subject cognitive workload decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 2324–2335, 2024.
[52] M. Jiménez-Guarneros, G. Fuentes-Pineda, and J. Grande-Barreto, "MMDA: A multimodal and multisource domain adaptation method for cross-subject emotion recognition from EEG and eye movement signals," *IEEE Trans. Comput. Social Syst.*, early access, 2024.

Minxu Liu received the B.S. degree in software engineering from Zhejiang University of Technology, Hangzhou, China, in 2023. He is currently working toward the M.S. degree in computer technology with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His research interests include brain decoding, affective computing, and domain adaptation.

Donghai Guan received the Ph.D. degree in computer engineering from Kyung Hee University (KHU), Suwon, Korea, in 2009.

He is an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He was a Research Professor with KHU during 2009–2011, Assistant Professor with KHU during 2012–2014. His research interests include machine learning, intelligent systems, and ubiquitous computing.

Chuhang Zheng received the B.S. degree in computer science and technology from Jimei University, Xiamen, China, in 2022. He is currently working toward the M.S. degree in computer technology with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His research interests include multimodal learning and affective computing.

Qi Zhu (Member, IEEE) received the B.S. and M.S. degrees in computer science and technology, and the Ph.D. degree in computer application technology from Harbin Institute of Technology, Harbin, China, in 2007, 2010, and 2014, respectively.

He is a Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He has published over 100 scientific articles in refereed international journal and conference proceedings. His research interests include pattern recognition, feature extraction, and medical image analysis.