Multimodal Emotion Recognition Using Deep Canonical Correlation Analysis

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, *Member, IEEE*, and Bao-Liang Lu, *Senior Member, IEEE*

Abstract—**Multimodal signals are more powerful than unimodal data for emotion recognition since they can represent emotions more comprehensively. In this paper, we introduce deep canonical correlation analysis (DCCA) to multimodal emotion recognition. The basic idea behind DCCA is to transform each modality separately and coordinate different modalities into a hyperspace by using specified canonical correlation analysis constraints. We evaluate the performance of DCCA on five multimodal datasets: the SEED, SEED-IV, SEED-V, DEAP, and DREAMER datasets. Our experimental results demonstrate that DCCA achieves state-of-the-art recognition accuracy rates on all five datasets: 94.58% on the SEED dataset, 87.45% on the SEED-IV dataset, 84.33% and 85.62% for two binary classification tasks and 88.51% for a four-category classification task on the DEAP dataset, 83.08% on the SEED-V dataset, and 88.99%, 90.57%, and 90.67% for three binary classification tasks on the DREAMER dataset. We also compare the noise robustness of DCCA with that of existing methods when adding various amounts of noise to the SEED-V dataset. The experimental results indicate that DCCA has greater robustness. By visualizing feature distributions with t-SNE and calculating the mutual information between different modalities before and after using DCCA, we find that the features transformed by DCCA from different modalities are more homogeneous and discriminative across emotions.**

Index Terms—**Multimodal signal, Multimodal emotion recognition, Multimodal deep learning, Deep canonical correlation analysis, EEG, Eye movement.**

I. INTRODUCTION

EMOTION strongly influences in our daily activities such as interactions between people, decision making, learning, and working. To endow a computer with emotion perception, understanding, and regulation abilities, Picard *et al.*

Wei Liu, and Bao-Liang Lu are with the Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, also with the Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: blu@cs.sjtu.edu.cn).

Jie-Lin Qiu is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Wei-Long Zheng is with the Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA.

developed the concept of affective computing, which aims to be used to study and develop systems and devices that can recognize, interpret, process, and simulate human affects [1], [2]. Human emotion recognition is a current hotspot in affective computing research. Since emotion recognition is critical for applications such as affective brain-computer interaction, emotion regulation and the diagnosis of emotion-related diseases, it is necessary to build a reliable and accurate model for recognizing human emotions.

Traditional emotion recognition systems are built with speech signals [3], facial expressions [4], and non-physiological signals [5]. However, in addition to clues from external appearances, emotions contain reactions from the central and peripheral nervous systems. Moreover, an obvious drawback of using behavioral modalities for emotion recognition is the uncertainty that arises in the case of individuals who either consciously regulate their emotional manifestations or are naturally suppressive. In contrast, EEG-based emotion recognition has been proven to be a reliable method because of its high recognition accuracy, objective evaluation and stable neural patterns [6], [7], [8], [9].

For the above reasons, researchers have tended to study emotions through physiological signals in recent years. These signals are more accurate and difficult to deliberately change by users. Lin and colleagues evaluated music-induced emotion recognition with EEG signals and attempted to use as few electrodes as possible [10]. Wang and colleagues used EEG signals to classify positive and negative emotions and compared different EEG features and classifiers [11]. Kim and André showed that electromyogram, electrocardiogram, skin conductivity, and respiration changes were reliable signals for emotion recognition [12]. Võ *et al.* studied the relationship between emotions and eye movement features, and they found that pupil diameters were influenced by both emotion and age [13].

Emotions are complex cognitive processes that involve subjective experience, expressive behaviors, and psychophysiological changes. Due to the rich characteristics of human emotions, it is difficult for single-modality signals to describe emotions comprehensively. Therefore, recognizing emotions with multiple modalities has become a promising method for building emotion recognition systems with high accuracy [14], [15], [16], [17], [18], [19]. Multimodal data can reflect emotional changes from multiple perspective, which is conducive to building a reliable and accurate emotion recognition model.

Multimodal fusion is one of the key aspects in taking full advantage of multimodal signals. In the past few years, researchers have utilized various methods to fuse different

modalities. Lu and colleagues employed feature-level concatenation, MAX fusion, SUM fusion, and fuzzy integral fusion to merge EEG and eye movement features, and they found the complementary properties of EEG and eye movement features in emotion recognition tasks [20]. Koelstra and colleagues evaluated the feature-level concatenation of EEG features and peripheral physiological features, and they found that participant ratings and EEG frequencies were significantly correlated and that decision fusion achieved the best emotion recognition results [21]. Sun *et al.* built a hierarchical classifier by combining both feature-level and decision-level fusion for emotion recognition tasks in the wild. The method was evaluated on several datasets and made very promising achievements on the validation and test sets [22].

Currently, with the rapid development of deep learning, researchers are applying deep learning models to fuse multiple modalities. Deep-learning-based multimodal representation frameworks can be classified into two categories: multimodal joint representation and multimodal coordinated representation [23]. Briefly, the multimodal joint representation framework takes all the modalities as input, and each modality starts with several individual neural layers followed by a hidden layer that projects the modalities into a joint space. The multimodal coordinated representation framework, instead of projecting the modalities together into a joint space, learns separate representations for each modality and coordinates them into a hyperspace with constraints between different modalities. Various multimodal joint representation frameworks have been applied to emotion recognition in very recent years [24], [25], [26], [27]. However, the multimodal coordinated representation framework has not yet been fully studied.

In this paper, we introduce a coordinated representation model named Deep Canonical Correlation Analysis (DCCA) [28], [29] to multimodal emotion recognition. The basic idea behind DCCA is to learn separate but coordinated representations for each modality under canonical correlation analysis (CCA) constraints. Since the coordinated representations are of the same dimension, we denote the coordinated hyperspace by $\mathcal{S}$.

Compared with our previous work [29], the main contributions of this paper on multimodal emotion recognition can be summarized as follows:

1. We introduce DCCA to multimodal emotion recognition and evaluate the effectiveness of DCCA on five benchmark datasets: the SEED, SEED-IV, SEED-V, DEAP, and DREAMER datasets. Our experimental results on these five datasets reveal that different emotions are disentangled in the coordinated hyperspace $\mathcal{S}$, and the transformation process of DCCA preserves emotion-related information and discards unrelated information.
2. We examine the robustness of DCCA and the existing methods on the SEED-V dataset under different levels of noise. The experimental results show that DCCA has higher robustness than the existing methods under most noise conditions.
3. By adjusting the weights of different modalities, DCCA allows users to fuse different modalities with greater flexibility such that various modalities contribute differently to the fused features.

The remainder of this paper is organized as follows. Section II summarizes the development and current state of multimodal fusion strategies. In Section III, we introduce the algorithms for the canonical correlation analysis, DCCA, the baseline models utilized in this paper, and the mutual information neural estimation (MINE) algorithm. The experimental settings are reported in Section IV. Section V presents and analyzes the experimental results. Finally, conclusions are given in Section VI.

II. Related Work

One of the key problems in multimodal deep learning is how to fuse data from different modalities. Multimodal fusion has gained increasing attention from researchers in diverse fields due to its potential for innumerable applications such as emotion recognition, event detection, image segmentation, and video classification [30], [31]. According to the level of fusion, traditional fusion strategies can be classified into the following three categories: 1) feature-level fusion (early fusion), 2) decision-level multimodal fusion (late fusion), and 3) hybrid multimodal fusion. With the rapid development of deep learning, an increasing number of researchers are employing deep learning models to facilitate multimodal fusion. In the following, we introduce these multimodal fusion types and their subtypes.

A. Feature-level fusion

Feature-level fusion is a common and straightforward method to fuse different modalities. The features extracted from the various modalities are first combined into a high-dimensional feature and then sent as a whole to the models [32], [21], [20], [33], [34].

The advantages of feature-level fusion are two-fold: 1) it can utilize the correlation between different modalities at an early stage, which better facilitates task accomplishment, and 2) the fused data contain more information than a single modality, and thus, a performance improvement is expected. The drawbacks of feature-level fusion methods mainly reside in the following: 1) it is difficult to represent the time synchronization between different modality features, 2) this type of fusion method might suffer the curse of dimensionality on small datasets, and 3) larger dimensional features might stress computational resources during model training.

B. Decision-level fusion

Decision-level fusion focuses on the usage of small classifiers and their combination. Ensemble learning is often used to assemble these classifiers. The term decision-level fusion describes a variety of methods designed to merge the outcomes and ensemble them into a single decision.

Rule-based fusion methods are most adopted in multimodal emotion recognition. Lu and colleagues utilized MAX fusion, SUM fusion, and fuzzy integral fusion for multimodal emotion recognition, and they found the complementary nature of

EEG and eye movement features by analyzing confusion matrices [20]. Although rule-based fusion methods are easy to use, the difficulty facing rule-based fusion is how to design good rules. If rules are too simple, they might not reveal the relationships between different modalities.

The advantage of decision-level fusion is that the decisions from different classifiers are easily compared and each modality can use its best suitable classifier for the task.

C. Hybrid fusion

Hybrid fusion is a combination of feature-level fusion and decision-level fusion. Sun and colleagues built a hierarchical classifier by combining both feature-level and decision-level fusion methods for emotion recognition [22]. Guo *et al.* built a hybrid classifier by combining fuzzy cognitive map and SVM to classify emotional states with compressed sensing representation [35].

D. Deep-learning-based fusion

For deep learning models, different types of multimodal fusion methods have been developed, and these methods can be grouped into two categories based on the modality representation: multimodal joint representation and multimodal coordinated representation [23].

The multimodal joint representation framework takes all the modalities as input, and each modality starts with several individual neural layers followed by a hidden layer that projects the modalities into a joint space. Both transformation and fusion processes are achieved automatically by black-box models and users do not know the meaning of the joint representations. The multimodal joint representation framework has been applied to emotion recognition [24], [25] and natural language processing [36].

The multimodal coordinated representation framework, instead of projecting the modalities together into a joint space, learns separate representations for each modality but coordinates them through a constraint. The most common coordinated representation models enforce similarity between modalities. Frome and colleagues proposed a deep visual semantic embedding (DeViSE) model to identify visual objects [37]. DeViSE is initialized from two pre-trained neural network models: a visual object categorization network and a skip-gram language model. DeViSE combines these two networks by the dot-product and hinge rank loss similarity metrics such that the model is trained to produce a higher dot product similarity between the visual model output and the vector representation of the correct label than that between the visual output and other randomly chosen text terms.

The deep canonical correlation analysis (DCCA) method, which is another model under the coordinated representation framework, was proposed by Andrew and colleagues [28]. In contrast to DeViSE, DCCA adopts traditional CCA as a similarity metric, which allows us to transform data into a highly correlated hyper-space.

III. METHODS

In this section, we first provide a brief description of traditional canonical correlation analysis (CCA) in Section III-A. Based on CCA, we present the building processes of DCCA in Section III-B. The baseline methods used in this paper are described in Section III-C. Finally, the mutual information neural estimation (MINE) algorithm is given in Section III-D, which is utilized to analyze the properties of transformed features implemented by DCCA in the coordinated hyperspace $\mathcal{S}$.

A. Canonical Correlation Analysis

Canonical correlation analysis (CCA) was proposed by Hotelling [38]. It is a widely used technique in the statistics community to measure the linear relationship between two multidimensional variables. Hardoon and colleagues applied CCA to machine learning [39].

Let $(X_1, X_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ denote random vectors with covariance matrices $(\Sigma_{11}, \Sigma_{22})$ and cross-variance matrix Σ_{12}. CCA attempts to find linear transformations of (X_1, X_2), $(w_1^* X_1, w_2^* X_2)$, which are maximally correlated:

$$(w_1^*, w_2^*) = \arg\max_{w_1, w_2} corr(w_1' X_1, w_2' X_2)$$
$$= \arg\max_{w_1, w_2} \frac{w_1' \Sigma_{12} w_2}{\sqrt{w_1' \Sigma_{11} w_1 w_2' \Sigma_{22} w_2}}. \qquad (1)$$

Since Eq. (1) is invariant to the scaling of the weights w_1 and w_2, Eq. (1) can be reformulated as follows:

$$(w_1^*, w_2^*) = \arg\max_{w_1' \Sigma_{11} w_1 = w_2' \Sigma_{22} w_2 = 1} w_1' \Sigma_{12} w_2, \qquad (2)$$

where we assume the projections are constrained to have unit variance.

To find multiple results of (w_1^i, w_2^i), subsequent projections are also constrained to be uncorrelated with previous projections, *i.e.*, $w_1^i \Sigma_{11} w_1^j = w_2^i \Sigma_{22} w_2^j = 0$ for $i < j$. Combining the top k projection vectors w_1^i into a matrix $A_1 \in \mathbb{R}^{n_1 \times k}$ as column vectors and similarly placing w_2^i into $A_2 \in \mathbb{R}^{n_2 \times k}$, we then identify the top $k \leq \min(n_1, n_2)$ projections:

$$\text{maximize:} \quad tr(A_1' \Sigma_{12} A_2)$$
$$\text{subject to:} \quad A_1' \Sigma_{11} A_1 = A_2' \Sigma_{22} A_2 = I. \qquad (3)$$

To solve this objective function, we first define $T = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$, and we let U_k and V_k be the matrices of the first k left singular and right singular vectors of T, respectively. Then the optimal objective value is the sum of the top k singular values of T, and the optimum is obtained at $(A_1^*, A_2^*) = (\Sigma_{11}^{-1/2} U_k, \Sigma_{22}^{-1/2} V_k)$. This method requires the covariance matrices Σ_{11} and Σ_{22} to be nonsingular, which is usually satisfied in practice.

For the original CCA, the representations in the latent space are obtained by linear transformations, which limit the scope of application of CCA. To address this problem, Lai and Fyfe [40] proposed kernel CCA, in which kernel methods are introduced for nonlinear transformations. Klami and colleagues developed probabilistic canonical correlation analysis (PCCA) [41]; later, they extended PCCA to a Bayesian-based

CCA named inter-battery factor analysis [42]. There are many other extensions of CCA such as tensor CCA [43], sparse CCA [44], and cluster CCA [45].

B. Deep Canonical Correlation Analysis

In this paper, we introduce deep canonical correlation analysis (DCCA) to multimodal emotion recognition. DCCA was proposed by Andrew and colleagues [28], and it computes representations of multiple modalities by passing them through multiple stacked layers of nonlinear transformations. Figure 1 depicts the structure of DCCA used in this paper.
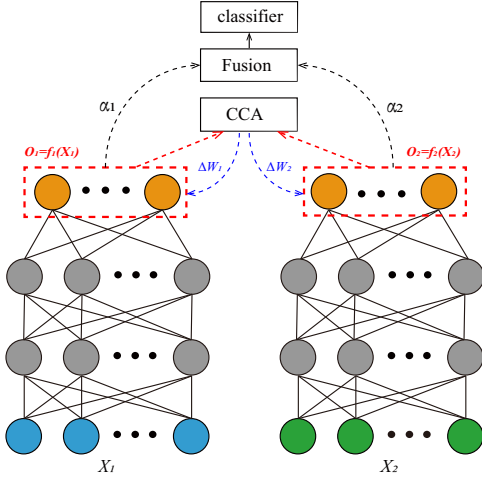

Fig. 1. The structure of DCCA. Different modalities are transformed by different neural networks separately. The outputs (O_1, O_2) are regularized by the traditional CCA constraint. Various strategies can be adopted to fuse O_1 and O_2, and we use the weighted sum fusion method as shown in Eq. (11). We update the parameters to maximize the CCA metric of different modalities, and the fused features are used to train a classifier.

Let $X_1 \in \mathbb{R}^{N \times d_1}$ be the instance matrix for the first modality and $X_2 \in \mathbb{R}^{N \times d_2}$ be the instance matrix for the second modality. Here, N is the number of instances, and d_1 and d_2 are the dimensions of the extracted features for these two modalities, respectively. To transform the raw features of two modalities nonlinearly, we build two deep neural networks for the two modalities as follows:

$$O_1 = f_1(X_1; W_1), \qquad (4)$$
$$O_2 = f_2(X_2; W_2), \qquad (5)$$

where W_1 and W_2 denote all parameters for the non-linear transformations, $O_1 \in \mathbb{R}^{N \times d}$ and $O_2 \in \mathbb{R}^{N \times d}$ are the outputs of the neural networks, and d denotes the output dimension of DCCA. The goal of DCCA is to jointly learn the parameters W_1 and W_2 for both neural networks such that the correlation of O_1 and O_2 is as high as possible:

$$(W_1^*, W_2^*) = \arg\max_{W_1, W_2} \; corr(f_1(X_1; W_1), F_2(X_2; W_2)). \quad (6)$$

We use the backpropagation algorithm to update W_1 and W_2. The solution to calculating the gradients of the objective function in Eq. (6) was developed by Andrew and colleagues [28]. Let $\bar{O}_1 = O_1' - \frac{1}{N} O_1' \mathbf{1}$ be the centered output matrix (similar to $\bar{O}_2$). We define $\hat{\Sigma}_{12} = \frac{1}{N-1} \bar{O}_1 \bar{O}_2'$,

$\hat{\Sigma}_{11} = \frac{1}{N-1} \bar{O}_1 \bar{O}_1' + r_1 \mathbf{I}$. Here, r_1 is a regularization constant (similar to $\hat{\Sigma}_{22}$). The total correlation of the top k components of O_1 and O_2 is the sum of the top k singular values of matrix $T = \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}$. In this paper, we take $k = d$, and the total correlation is the trace of T:

$$corr(O_1, O_2) = \left(tr(T'T) \right)^{1/2}. \qquad (7)$$

Finally, we calculate the gradients with the singular decomposition of $T = UDV'$,

$$\frac{\partial corr(O_1, O_2)}{\partial O_1} = \frac{1}{N-1}(2\nabla_{11}\bar{O}_1 + \nabla_{12}\bar{O}_2), \qquad (8)$$

where

$$\nabla_{11} = -\frac{1}{2}\hat{\Sigma}_{11}^{-1/2} U D U' \hat{\Sigma}_{11}^{-1/2}, \qquad (9)$$
$$\nabla_{12} = \hat{\Sigma}_{11}^{-1/2} U V' \hat{\Sigma}_{22}^{-1/2}, \qquad (10)$$

and $\partial corr(O_1, O_2)/\partial O_2$ has a symmetric expression.

After the training of the two neural networks, the transformed features $O_1, O_2 \in \mathcal{S}$ are in the coordinated hyperspace $\mathcal{S}$. In the original DCCA [28], the authors did not explicitly describe how to use transformed features for real-world applications via machine learning algorithms. Users need to design a strategy to take advantage of the transformed features according to their application.

In this paper, we use a weighted sum fusion method to obtain the fused features as follows:

$$O = \alpha_1 O_1 + \alpha_2 O_2, \qquad (11)$$

where α_1 and α_2 are weights satisfying $\alpha_1 + \alpha_2 = 1$. The fused features O are used to train the classifiers to recognize different emotions. In this paper, an SVM classifier is adopted.

According to the construction processes mentioned above, DCCA brings the following advantages to multimodal emotion recognition:

- By transforming different modalities separately, we can explicitly extract transformed features for each modality (O_1 and O_2) so that it is convenient to examine the characteristics and relationships of modality-specific transformations.
- With specified CCA constraints, we can regulate the nonlinear mappings ($f_1(\cdot)$ and $f_2(\cdot)$) and make the model preserve the emotion-related information.
- By using a weighted sum fusion (under the condition $\alpha_1 + \alpha_2 = 1$), we can assign different priorities to these modalities based on our priori knowledge. A larger weight represents a larger contribution of the corresponding modality to the fusion features.

C. Baseline methods

1) Concatenation Fusion: The concatenation fusion is a type of feature-level fusion. The feature vectors from two modalities are denoted as $X^1 = [x_1^1, \cdots, x_n^1] \in \mathcal{R}^n$ and $X^2 = [x_1^2, \cdots, x_m^2] \in \mathcal{R}^m$, and the fused features can be calculated with the following equation:

$$\begin{aligned} X_{fusion} &= Concat([X^1, X^2]) \\ &= [x_1^1, \cdots, x_n^1, x_1^2, \cdots, x_m^2]. \end{aligned} \qquad (12)$$

2) MAX Fusion: The MAX fusion method is a type of decision-level fusion method that chooses the class of the maximum probability as the prediction result. Assuming that we have K classifiers and C categories, there is a probability distribution for each sample $P_j(Y_i|x_t), j \in \{1, \cdots, K\}$, and $i \in \{1, \cdots, C\}$, where x_t is a sample, Y_i is the predicted label, and $P_j(Y_i|x_t)$ is the probability of sample x_t belonging to class i generated by the j-th classifier. The MAX fusion rule can be expressed as follows:

$$\hat{Y} = \arg\max_i\{\arg\max_j P_j(Y_i|x_t)\}. \tag{13}$$

3) Fuzzy Integral Fusion: The fuzzy integral fusion is also a type of decision-level fusion [46], [47]. A fuzzy measure μ on the set X is a function: $\mu : \mathcal{P}(X) \rightarrow [0,1]$, which satisfies the two axioms: 1) $\mu(\emptyset) = 0$ and 2) $A \subset B \subset X$ implies $\mu(A) \leq \mu(B)$. In this paper, we use the discrete Choquet integral to fuse the multimodal features. The discrete Choquet integral of a function $f : X \rightarrow \mathcal{R}^+$ with respect to μ is defined by

$$\mathcal{C}_\mu(f) := \sum_{i=1}^{n} \big(f(x_i) - f(x_{i-1})\big)\mu(A_{(i)}), \tag{14}$$

where $\cdot_{(i)}$ indicates that the indices have been permuted such that $0 \leq f(x_{(1)}) \leq \cdots \leq f(x_{(n)})$, $A_{(i)} := \{x_{(i)}, \cdots, x_{(n)}\}$, and $f(x_{(0)}) = 0$.

In this paper, we utilize the algorithm proposed by Tanaka and Sugeno [48] to calculate the fuzzy measure. The algorithm attempts to find the fuzzy measure μ, which minimizes the total squared error of the model. Tanaka and Sugeno proved that the minimization problem can be solved through a quadratic programming method.

4) Bimodal Deep AutoEncoder (BDAE): BDAE was proposed by Ngiam and colleagues [33]. In our previous work, we applied BDAE to multimodal emotion recognition [24].

A building block of BDAE is the restricted Boltzmann machine (RBM). The RBM is an undirected graph model, which has a visible layer and a hidden layer. Connections exist only between the visible layer and hidden layer, and there are no connections in the visible layer or in the hidden layer. In this paper, we adopted the *BernoulliRBM* in Scikit-learn[1] [49]. The visible variables $\mathbf{v} \sim \text{Bern}(p)$ are binary stochastic units of dimension M, which means that the input data should be either binary or real valued between 0 and 1, signifying the probability. The hidden variables also satisfy a Bernoulli distribution $\mathbf{h} \in \{0,1\}^N$. The energy is calculated with the following function:

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{M}\sum_{j=1}^{N} W_{ij}v_ih_j - \sum_{i=1}^{M} b_iv_i - \sum_{j=1}^{N} a_jh_j, \tag{15}$$

where $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$ are parameters, W_{ij} is the symmetric weight between the visible unit i and the hidden unit j, and b_i and a_j are the bias terms of the visible unit and hidden unit,

[1]https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.BernoulliRBM.html

respectively. With an energy function, we can obtain the joint distribution over the visible and hidden units:

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{\mathcal{Z}(\theta)} \exp(E(\mathbf{v}, \mathbf{h}; \theta)), \tag{16}$$

$$\mathcal{Z}(\theta) = \sum_{\mathbf{v}}\sum_{\mathbf{h}} \exp(E(\mathbf{v}, \mathbf{h}; \theta)), \tag{17}$$

where $\mathcal{Z}(\theta)$ is the normalization constant. Given a set of visible variables $\{\mathbf{v}_n\}_{n=1}^N$, the derivative of the log-likelihood with respect to the weight $\mathbf{W}$ can be calculated from Eq. (16):

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \log p(\mathbf{v}_n; \theta)}{\partial W_{ij}} = \mathbb{E}_{P_{data}}[v_ih_j] - \mathbb{E}_{P_{model}}[v_ih_j]. \tag{18}$$

The BDAE training procedure includes encoding and decoding. In the encoding phase, we train two RBMs for EEG features and eye movement features, and the hidden layers are denoted as h_{EEG} and h_{Eye}. These two hidden layers are concatenated together, and the concatenated layer is used as the visual layer of a new upper RBM. In the decoding stage, we unfold the stacked RBMs to reconstruct the input features. Finally, we use a back-propagation algorithm to minimize the reconstruction error.

D. Mutual Information Neural Estimation

Mutual information is a fundamental quantity for measuring the relationship between variables. The mutual information quantifies the dependence of two random variables X and Z with the following equation:

$$I(X; Z) = \int_{\mathcal{X} \times \mathcal{Z}} \log \frac{d\mathbb{P}_{XZ}}{d\mathbb{P}_X \otimes \mathbb{P}_Z} d\mathbb{P}_{XZ}, \tag{19}$$

where $\mathbb{P}_{XZ}$ is the joint probability distribution, and $\mathbb{P}_X = \int_Z d\mathbb{P}_{XZ}$ and $\mathbb{P}_Z = \int_X d\mathbb{P}_{XZ}$ are the marginals.

The mutual information neural estimation (MINE) was proposed by Belghazi and colleagues [50]. MINE is linearly scalable in dimensionality as well as in sample size, trainable through a back-propagation algorithm, and strongly consistent.

The idea behind MINE is to choose $\mathcal{F}$ to be the family of functions $T_\theta : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}$ parameterized by a deep neural network with parameters $\theta \in \Theta$. Then, the deep neural network is used to update the estimated mutual information,

$$I(X; Z) \geq I_\Theta(X; Z), \tag{20}$$

where I_Θ is defined as

$$I_\Theta(X; Z) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^{T_\theta}]). \tag{21}$$

The expectations in Eq. (21) are estimated using empirical samples from $\mathbb{P}_{XZ}$ and $\mathbb{P}_X \otimes \mathbb{P}_Z$ or by shuffling the samples from the joint distribution, and MINE is defined as

$$\widehat{I(X; Z)}_n = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}^{(n)}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X^{(n)} \otimes \hat{\mathbb{P}}_Z^{(n)}}[e^{T_\theta}]), \tag{22}$$

where $\hat{\mathbb{P}}$ is the empirical distribution associated with n *i.i.d.* samples. The details on the implementation of MINE are provided in Algorithm 1.

Algorithm 1 Mutual Information Calculation between Two Modalities with MINE

Input: Features extracted from two modalities:
$$X = \{x^1, \cdots, x^n\}, Z = \{z^1, \cdots, z^n\}$$
Output: Estimated mutual information
$\theta \leftarrow$ initialize network parameters
repeat
  1. Draw b mini-batch samples from the joint distribution:
    $(x^{(1)}, z^{(1)}), \ldots, (x^{(b)}, z^{(b)}) \sim \mathbb{P}_{XZ}$
  2. Draw b samples from the Z marginal distribution:
    $\bar{z}^{(1)}, \ldots, \bar{z}^{(b)} \sim \mathbb{P}_Z$
  3. Evaluate the lower bound:
    $\mathcal{V}(\theta) \leftarrow \frac{1}{b} \sum T_\theta(x^{(i)}, z^{(i)}) - log(\frac{1}{b} \sum e^{T_\theta(x^{(i)}, z^{(i)})})$
  4. Evaluate bias-corrected gradients:
    $\hat{G}(\theta) \leftarrow \tilde{\nabla}_\theta \mathcal{V}(\theta)$
  5. Update the deep neural network parameters:
    $\theta \leftarrow \theta + \hat{G}(\theta)$
until Convergence

We modify the code of the MINE algorithm written by Masanori Yamada[2]; the code used in this paper can be downloaded from GitHub[3].

IV. EXPERIMENTAL SETTINGS

A. Datasets

To evaluate the effectiveness of DCCA for multimodal emotion recognition, five multimodal emotion recognition datasets are selected for experimental study in this paper.

1) SEED dataset[4]: The SEED dataset was developed by Zheng and Lu [6]. A total of 15 Chinese film clips of three emotions (happy, neutral and sad) were chosen from a pool of materials as stimuli used in the experiments. Before the experiments, the participants were told the procedures of the entire experiment. During the experiments, the participants were asked to watch the selected 15 movie clips, and report their emotional feelings. After watching a movie clip, the subjects were given 45 seconds to provide feedback and 15 seconds to rest. In this paper, we use the same subset of the SEED dataset as in our previous work [20], [24], [25] for the comparison study.

The SEED dataset contains EEG signals and eye movement signals. The EEG signals were collected with an ESI NeuroScan system at a sampling rate of 1000 Hz from a 62-channel electrode cap. Eye movement signals were collected with SMI eye tracking glasses[5].

2) SEED-IV dataset: The SEED-IV dataset was first proposed in [15]. The experimental procedure was similar to that of the SEED dataset, and 72 film clips were chosen as stimuli materials. The dataset contains emotional EEG signals and eye movement signals of four different emotions, *i.e.*, happy, sad, neutral, and fear. A total of 15 subjects (7 male and 8 female) participated in the experiments. For each participant,

three sessions were performed on different days, and each session consisted of 24 trials. In each trial, the participant watched one of the movie clips.

3) SEED-V dataset: The SEED-V dataset was proposed in [51]. The dataset contains EEG signals and eye movement signals for five emotions (happy, sad, neutral, fear, and disgust). A total of 16 subjects (6 male and 10 female) were recruited to participate in the experiment, and each of them performed the experiment three times. During the experiment, the subject were required to watch 15 movie clips (3 clips for each emotion). The same devices were used in the SEED-V dataset as in the SEED and SEED-IV datasets. The SEED-V dataset used in this paper will be freely available to the academic community as a subset of SEED[6].

4) DEAP dataset: The DEAP dataset was developed by Koelstra and colleagues [21] and is a multimodal dataset for the analysis of human affective states. The EEG signals and peripheral physiological signals (EOG, EMG, GSR, respiration belt, and plethysmograph) of 32 participants were recorded as each watched 40 one-minute long excerpts of music videos. Participants rated each video in terms of the levels of arousal, valence, like/dislike, dominance, and familiarity.

5) DREAMER dataset: The DREAMER dataset is a multimodal emotion dataset developed by Katsigiannis and Ramzan [52]. The DREAMER dataset consists of 14-channel EEG signals and 2-channel ECG signals of 23 subjects (14 males and 9 females). During the experiments, the participants watched 18 film clips to elicit 9 different emotions including amusement, excitement, happiness, calmness, anger, disgust, fear, sadness, and surprise. After watching a clip, the self-assessment manikins were used to acquire subjective assessments of valence, arousal, and dominance.

B. Feature extraction

1) EEG feature extraction: For EEG signals, we extract differential entropy (DE) features using short-term Fourier transforms with a 4-second Hanning window without overlapping [53], [54]. The differential entropy feature is used to measure the complexity of continuous random variables. Its calculation formula can be written as follows:

$$h(\mathbf{X}) = - \int_\mathbf{X} f(x) \log(f(x)) dx, \tag{23}$$

where $\mathbf{X}$ is a random variable and $f(x)$ is the probability density function of $\mathbf{X}$. For the time series $\mathbf{X}$, obeying the Gauss distribution $N(\mu, \sigma^2)$, its differential entropy can be calculated as follows:

$$h(\mathbf{X}) = - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx$$
$$= \frac{1}{2} \log 2\pi e \sigma^2. \tag{24}$$

Shi and colleagues [54] proved that EEG signals within a short time period in different frequency bands are subject to a Gaussian distribution by the Kolmogorov-Smirnov test, and the DE features can be calculated by Eq. (24).

---

[2]https://github.com/MasanoriYamada/Mine_pytorch

[3]https://github.com/csliuwei/MI_plot

[4]http://bcmi.sjtu.edu.cn/home/seed/index.html

[5]https://www.smivision.com/eye-tracking/product/eye-tracking-glasses/

[6]http://bcmi.sjtu.edu.cn/home/seed/index.html

We extract DE features from EEG signals (from the SEED, SEED-IV and SEED-V datasets) in five frequency bands for all channels: delta (1-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-31 Hz), and gamma (31-50 Hz). There are in total $62 \times 5 = 310$ dimensions for 62 EEG channels. Finally we adopt the linear dynamic system method to filter out noise and artifacts [55].

For the DEAP dataset, the raw EEG signals were downsampled to 128 Hz and preprocessed with a bandpass filter from 4 to 75 Hz. We extract the DE features from four frequency bands (theta, alpha, beta, and gamma). As a result, there are 128 dimensions for the DE features.

2) ECG feature extraction: In previous ECG-based emotion recognition studies, researchers extracted time-domain features, frequency-domain features, and time-frequency-domain features from ECG signals for emotion recognition [52], [56], [57]. Katsigiannis and Ramzan extracted power spectral density (PSD) features of low frequency and high frequency from ECG signals [52]. Hsu and colleagues extracted power for three frequency bands: a very-low-frequency range (0.0033 – 0.04 Hz), a low-frequency range (0.04 – 0.15 Hz), and a high-frequency range (0.15 – 0.4 Hz) [56].

However, previous studies have shown that ECG signals have a much wider frequency range. In the early stage of ECG research, Scher and Young showed that ECG signals contain frequency components as high as 100 Hz [58]. Recently, Shufni and Mashor also showed that there are high-frequency components (up to 600 Hz) in ECG signals [59]. Tereshchenko and Josephson reviewed studies on ECG frequencies and noted that "the full spectrum of frequencies producing the QRS complex has not been adequately explored" [60].

Since there are no standard frequency separation methods for ECG signals [60], we extract the logarithm of the average energy of five frequency bands (1 – 4 Hz, 4 – 8 Hz, 8 – 14 Hz, 14 – 31 Hz, and 31 – 50 Hz) from two ECG channels of the DREAMER dataset. As a result, we extract 10-dimensional features from the ECG signals.

3) Eye movement features: The eye movement data in the SEED dataset recorded using SMI ETG eye-tracking glasses[5] provide various types of parameters such as pupil diameters, fixation positions and durations, saccade information, blink details, and other event statistics. Although emotional changes cause fluctuations in pupil diameter, environmental luminance is the main reason for pupil diameter changes. Consequently, we adopt a principal component analysis-based method to remove the changes caused by lighting conditions [16].

The eye movement signals acquired by SMI ETG eye-tracking glasses contain both statistical features, such as blink information, and computational features such as temporal and frequency features. Table I shows all 33 eye movement features used in this paper. Therefore, the total number of dimensions of the eye movement features is 33.

4) Peripheral physiological signal features: For peripheral physiological signals from the DEAP dataset, we calculate statistical features in the temporal domain, including the maximum value, minimum value, mean value, standard deviation, variance, and squared sum. Since there are 8 channels for

TABLE I
SUMMARY OF EXTRACTED EYE MOVEMENT FEATURES.

Eye movement parameters	Extracted features
Pupil diameter (X and Y)	Mean, standard deviation, DE in four bands (0–0.2Hz,0.2–0.4Hz, 0.4–0.6Hz,0.6–1Hz)
Disperson (X and Y)	Mean, standard deviation
Fixation duration (ms)	Mean, standard deviation
Blink duration (ms)	Mean, standard deviation
Saccade	Mean and standard deviation of saccade duration(ms) and saccade amplitude(°)
Event statistics	Blink frequency, fixation frequency, fixation duration maximum, fixation dispersion total, fixation dispersion maximum, saccade frequency, saccade duration average, saccade amplitude average, saccade latency average.

the peripheral physiological signals, we extract 48 (6×8)-dimensional features.

C. Model training

For the SEED dataset, the DE features of the first 9 movie clips are used as training data, and those of the remaining 6 movie clips are used as test data. In this paper, we build subject-dependent models to classify three types of emotions (happy, sad, and neutral), which is the same as in our previous work [20], [24], [25].

A similar training-testing separation scheme is applied to the SEED-IV dataset. There are 24 trials for each session, and we use the data from the first 16 trials as the training data and the data from the remaining 8 trials as the test data [15]. DCCA is trained to recognize four emotions (happy, sad, fear, and neutral)

For the SEED-V dataset, the training-testing separation strategy is the same as that used by Zhao *et .al* [61]. We adopt three-fold cross-validation to evaluate the performance of DCCA on five emotion (happy, sad, fear, neutral, and disgust) recognition tasks. Since the participant watched 15 movie clips (the first 5 clips, the middle 5 clips and the last 5 clips) and participated in three sessions, we concatenate features of the first 5 clips from three sessions (*i.e.*, we concatenate features extracted from 15 movie clips) as the training data for fold one (with a similar operation for folds two and three).

For the DEAP dataset, we build a subject-dependent model with a 10-fold cross-validation on two binary classification tasks and a four-emotion recognition task:

- Binary classifications: arousal-level and valence-level classification with a threshold of 5.
- Four-category classification: high arousal, high valence (HAHV); high arousal, low valence (HALV); low arousal, high valence (LAHV); and low arousal, low valence (LALV).

For the DREAMER dataset, we utilize leave-one-out cross-validation (*i.e.*, 18-fold validation) to evaluate the performance

TABLE II
SUMMARY OF THE DCCA STRUCTURES FOR FIVE DIFFERENT DATASETS

Datasets	#Hidden Layers	#Hidden Units	Output Dimensions
SEED	6	400±40, 200±20, 150±20, 120±10, 60±10, 20±2	20
SEED-IV	7	400±40, 200±20, 150±20, 120±10, 90±10, 60±10, 20±2	20
SEED-V	2	searching for the best numbers between 50 and 200	12
DEAP	7	1500±50, 750±50, 500±25, 375±25, 130±20, 65±20, 30±20	20
DREAMER	2	searching for the best numbers between 10 and 200	5

of DCCA on three binary classification tasks (arousal, valence, and dominance), which is the same as that used by Song *et al.* [62].

For these five different datasets, DCCA uses different hidden layers, hidden units, and output dimensions. Table II summarizes the DCCA structures for these datasets. For all five datasets, the learning rate, batch size, and regulation parameter of DCCA are set to 0.001, 100, and $1e^{-8}$, respectively.

V. EXPERIMENTAL RESULTS

A. SEED, SEED-IV, and DEAP Datasets

In this section, we summarize our previous results on SEED, SEED-IV, and DEAP datasets [29]. Table III lists the results obtained by seven existing methods and DCCA on the SEED dataset. Lu and colleagues applied concatenation fusion, MAX fusion and fuzzy integral to fuse multiple modalities and demonstrated that the fuzzy integral fusion method achieved the accuracy of 87.59% [20]. Liu *et al.* [24] and Tang *et al.* [25] improved multimodal methods, obtaining accuracies of 91.01% and 94.58%, respectively. Recently, Yang and colleagues [8] build a single-layer feedforward network (SLFN) with subnetwork nodes and achieved an accuracy of 91.51%. Song and colleagues [62] proposed DGCNN and obtained a classification accuracy of 90.40%. As seen from Table III, DCCA achieves the best result of 94.58% among the eight different methods.

TABLE III
THE MEAN ACCURACY RATES (%) AND STANDARD DEVIATIONS (%) OF
SEVEN EXISTING METHODS AND DCCA ON THE SEED DATASET

Methods	Mean	Std
Concatenation [20]	83.70	-
MAX [20]	81.71	-
FuzzyIntegral [20]	87.59	19.87
BDAE [24]	91.01	8.91
DGCNN [62]	90.40	8.49
SLFN with subnetwork nodes [8]	91.51	–
Bimodal-LSTM [25]	93.97	7.03
DCCA	**94.58**	**6.16**

Table IV gives the results of five different methods on the SEED-IV dataset. We can observe from Table IV that for the SVM classifier, the four emotion states are recognized with a 75.88% mean accuracy rate, and the BDAE model improved the result to 85.11%. DCCA outperforms the aforementioned two methods, with an 87.45% mean accuracy rate.

Two classification schemes are adopted on the DEAP dataset. Table V shows the results of binary classifications. As we can observe, DCCA achieves the best results in both arousal classification (84.33%) and valence classification (85.62%) tasks.

TABLE IV
THE MEAN ACCURACY RATES (%) AND STANDARD DEVIATIONS (%) OF
FOUR EXISTING METHODS AND DCCA ON THE SEED-IV DATASET

Methods	Mean	Std
Concatenation	77.63	16.43
MAX	68.99	17.14
FuzzyIntegral	73.55	16.72
BDAE [15]	85.11	11.79
DCCA	**87.45**	**9.23**

For the four-category classification task on the DEAP dataset, Zheng and colleagues [7] adopted the GELM model and achieved an accuracy of 69.67%. Chen *et al.* [63] proposed a three-stage decision framework that outperformed KNN and SVM with an accuracy rate of 70.04%. The DCCA model achieved a mean accuracy rate of 88.51%, which is more than 18% higher than the existing methods.

TABLE V
THE MEAN ACCURACY RATES (%) AND STANDARD DEVIATION (%) OF
THREE EXISTING METHODS AND DCCA FOR THE TWO BINARY EMOTION
CLASSIFICATION TASKS ON THE DEAP DATASET.

Methods	Arousal	Valence
BDAE [24]	80.50/3.39	85.20/4.47
MESAE [27]	84.18/-	83.04/-
Bimodal-LSTM [25]	83.23/2.61	83.82/5.01
DCCA	**84.33/2.25**	**85.62/3.48**

TABLE VI
THE MEAN ACCURACY RATES (%) AND STANDARD DEVIATIONS (%) OF
TWO EXISTING METHODS AND DCCA FOR THE FOUR-CATEGORY
CLASSIFICATION TASK ON THE DEAP DATASET.

Methods	Acc
Three-stage decision Framework [63]	70.04/-
GELM [7]	69.67/-
DCCA	**88.51/8.52**

From the experimental results mentioned above, we can see that DCCA outperforms the existing methods on the SEED, SEED-IV, and DEAP datasets.

B. SEED-V Dataset

We examine the effectiveness of DCCA on the SEED-V dataset, which contains multimodal signals of five emotions (happy, sad, fear, neutral, and disgust).

We perform a series of experiments to choose the best output dimension and fusion coefficients (α_1 and α_2 in Eq. (11)) for DCCA. We adopt the grid search method with output dimensions ranging from 5 to 50 and coefficients for the EEG features ranging from 0 to 1, *i.e.* $\alpha_1 = [0, 0.1, 0.2, \cdots, 0.9, 1.0]$.

Since $\alpha_1 + \alpha_2 = 1$, we can calculate the weight for the other modality via $\alpha_2 = 1 - \alpha_1$. Figure 2 shows the heat map of the experimental results of the grid search. Each row in Fig. 2 gives different output dimensions, and each column is the weight of the EEG features (α_1). The numbers in blocks are the accuracy rates, which are rounded to integers for simplicity. According to Fig. 2, we set the output dimension to 12 and the weight of the EEG features to 0.7 (*i.e.*, $\alpha_1 = 0.7, \alpha_2 = 0.3$).
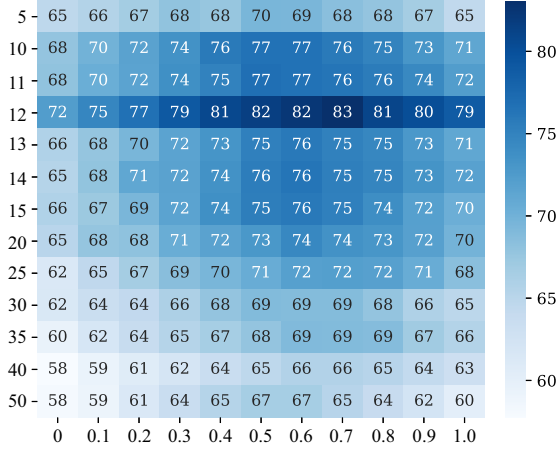
Fig. 2. Selection of the best output dimension and EEG weight of DCCA on the SEED-V dataset. Each row represents the number of output dimensions, and each column denotes the weight (α_1) of the EEG features.
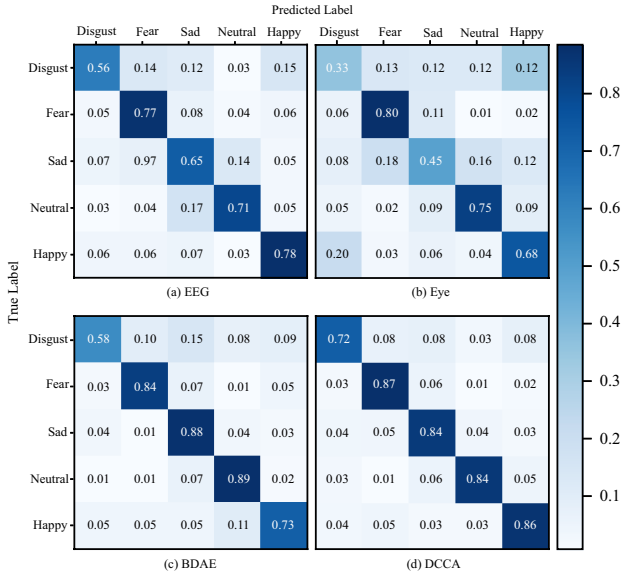

Fig. 3. Comparison of the confusion matrices of different methods on the SEED-V dataset. Subfigures (a), (b), and (c) are the confusion matrices from [61] for SVM classifiers of unimodal features and BDAE model of multimodal features. Subfigure (d) is the confusion matrix of DCCA.

Table VII summarizes the emotion recognition results on the SEED-V dataset. Zhao and colleagues [61] adopted feature-level concatenation and the bimodal deep autoencoder (BDAE) for fusing multiple modalities, and achieved mean accuracy rates of 73.65% and 79.70%, respectively. In addition to feature-level concatenation, we also implement MAX fusion and fuzzy integral fusion strategies here. As shown in Table

VII, the MAX fusion and fuzzy integral fusion yielded mean accuracy rates of 73.14% and 73.62%, respectively. The mean accuracy rate of DCCA is 83.08%, which is the best result among the five fusion strategies.

TABLE VII
THE MEAN ACCURACY RATES (%) AND STANDARD DEVIATIONS (%) OF FOUR EXISTING METHODS AND DCCA ON THE SEED-V DATASET

Methods	Mean	Std
Concatenation [61]	73.65	8.90
MAX	73.17	9.27
FuzzyIntegral	73.24	8.72
BDAE [61]	79.70	**4.76**
DCCA	**83.08**	7.11

Figure 3 depicts the confusion matrices of the DCCA model and the models adopted by Zhao and colleagues [61]. Figures. 3(a), (b) and (c) are the confusion matrices for the EEG features, eye movement features, and the BDAE model, respectively. Figure 3(d) depicts the confusion matrix for the DCCA model. From Figs. 3(a), (b), and (d), for each of the five emotions, DCCA achieves a higher accuracy, indicating that emotions are better represented and more easily classified in the coordinated hyperspace $\mathcal{S}$ transformed by DCCA.

From Figs. 3(a) and (c), compared with the unimodal results of the EEG features, the BDAE model achieves worse classification results on the happy emotion, suggesting that the BDAE model might not take full advantage of different modalities for the happy emotion. Comparing Figs. 3(c) and (d), DCCA largely improved the classification results on disgust and happy emotion recognition tasks compared with the BDAE model, implying that DCCA is more effective in fusing multiple modalities.

To analyze the coordinated hyperspace $\mathcal{S}$ of DCCA, we utilized the t-SNE algorithm to visualize the space of the original features and the coordinated hyperspace of the transformed features and fused features. Figure 4 presents a visualization of the features from three participants. The first row shows the original features, the second row depicts the transformed features, and the last row presents the fused features. The different colors stand for different emotions, and the different markers are different modalities. We can make the following observations:

- Different emotions are disentangled in the coordinated hyperspace $\mathcal{S}$. For original features, there are more overlaps among different emotions (different colors presenting substantial overlap), which lead to poorer emotional representation. After the DCCA transformation, different emotions become relatively independent, and the overlapping areas are considerably reduced. This indicates that the transformed features have improved emotional representation capabilities compared with the original features. Finally, after multimodal fusion, different emotions ('□' of different colors in the last row) are completely separated, and there is no overlapping area, indicating that the merged features also have good emotional representation ability.
- Different modalities have homogeneous distributions in the coordinated hyperspace $\mathcal{S}$. To make this observation
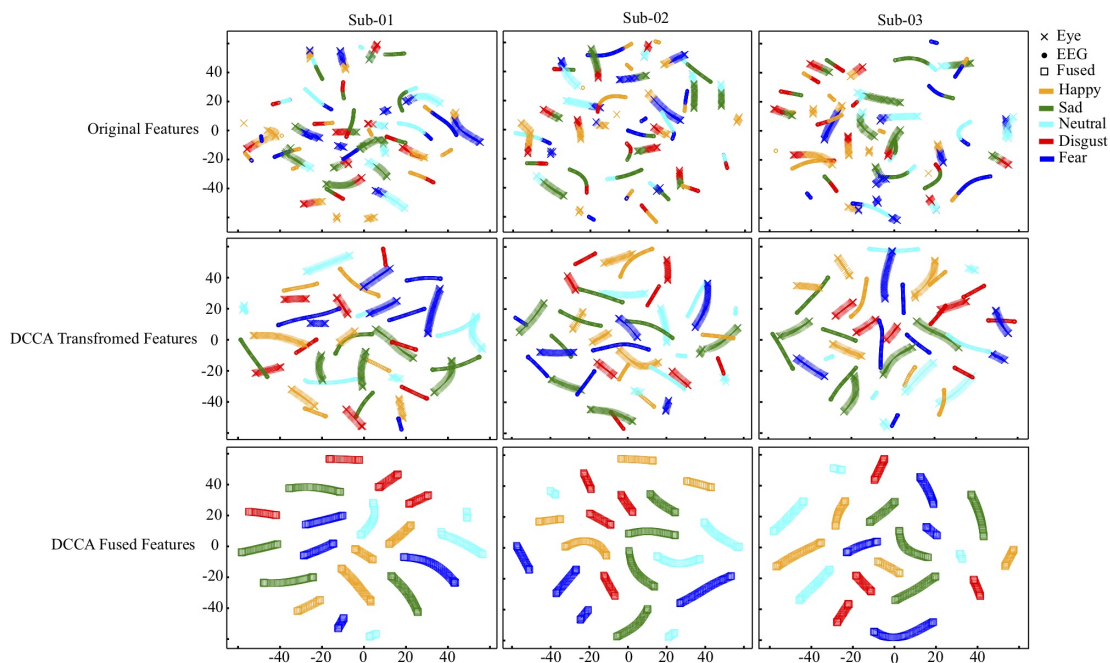
Fig. 4. Feature distribution visualization by the t-SNE algorithm. The original features, transformed features, and fused features from the three subjects are presented. The different colors stand for different emotions, and the different markers indicate different features.
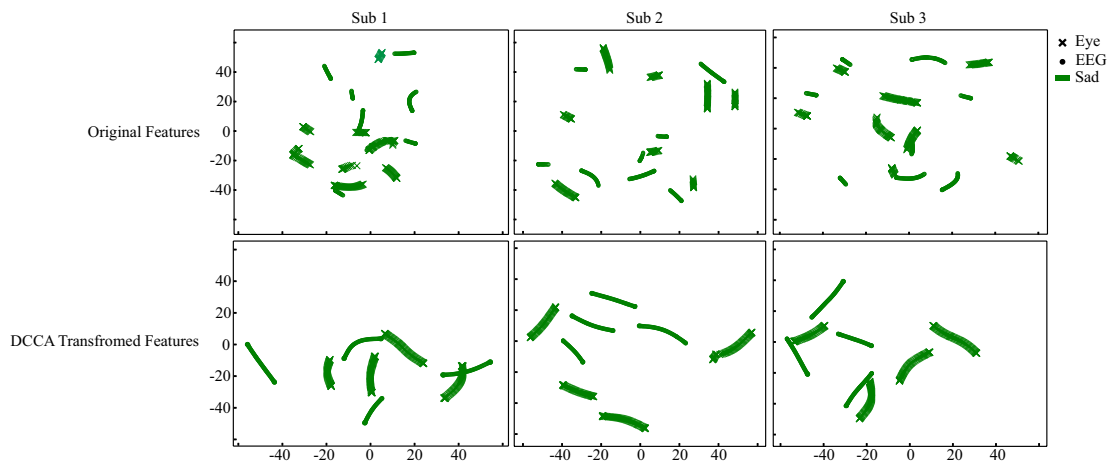

Fig. 5. Distributions of EEG and eye movement features for the sad emotion. The transformed features have more compact distributions from both inter-modality and intra-modality perspectives.

more obvious, we separate and plot the distributions of the EEG and eye movement features under the sad emotion in Fig. 5. From the perspectives of both inter-modality and intra-modality distributions, the original EEG features ('○' marker) and eye movement features ('×' marker) are separated from each other. After the DCCA transformation, the EEG features and the eye movement features have more compact distributions, indicating that the coordinated hyperspace preserves shared emotion-related information and discards irrelevant information.

Figures 4 and 5 qualitatively show that DCCA maps original EEG and eye movement features into a coordinated hyperspace S where emotions are better represented since only emotion related information is preserved.

Furthermore, we calculated the mutual information of the original features and transformed features to support our claims quantitatively. Figure 6 presents the mutual information of three participants estimated by MINE. The green curves depict the mutual information of the original EEG and eye movement features, and the red curves are the estimated mutual information of the transformed features. The transformed features have more mutual information than the original features, indicating that EEG and eye movement features in the coordinated hyperspace provide more shared emotion-related information, which is consistent with observations from Figs. 4 and 5.
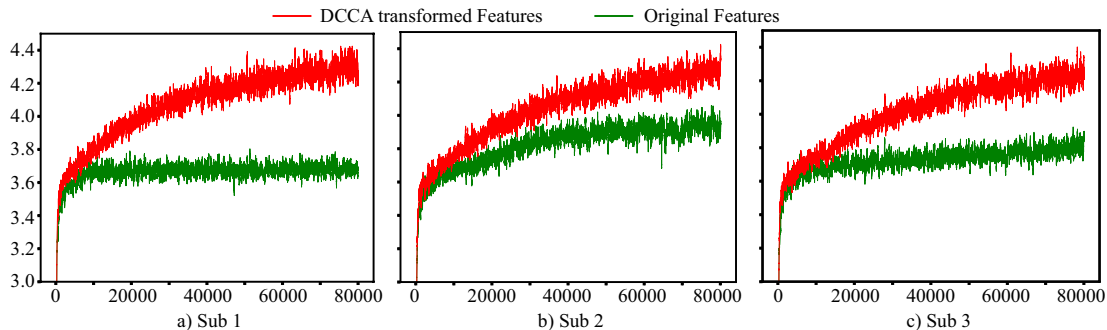
Fig. 6. Mutual information (MI) estimation with MINE. The green curve shows the estimated MI for the original EEG features and eye movement features. The red curve depicts the MI for the transformed features. The x axis is the epoch number of the deep neural network used to estimate MI, and the y axis is the estimated MI. Moving average smoothing is used to smooth the curves.

C. Robustness Analysis on the SEED-V Dataset

EEG signals have a low signal-to-noise ratio (SNR) and are easily interfered with by external environmental noise. To compare the noise robustness of DCCA with that of the existing methods, we designed two experimental schemes on noisy datasets: 1) we added Gaussian noise of different variances to both the EEG and eye movement features. To highlight the influence of noise, we added noise to the normalized features since the directly extracted features are much larger than the generated noise (which is mostly less than 1). 2) Under certain extreme conditions, EEG signals may be overwhelmed by noise. To simulate this situation, we randomly replace different proportions (10%, 30%, and 50%) of EEG features with noise with a normal distribution ($X \sim \mathcal{N}(0,1)$), gamma distribution ($X \sim \Gamma(1,1)$), and uniform distribution ($X \sim \mathcal{U}[0,1]$). Specifically, for DCCA, we also examine the effect of different weight coefficients on the robustness of the model. In this paper, we compare the performance of three different combinations of coefficients, *i.e.*, $\alpha_1 = 0.3$ (DCCA-0.3), $\alpha_1 = 0.5$ (DCCA-0.5), and $\alpha_1 = 0.7$ (DCCA-0.7).

1) Adding Gaussian noise: First, we investigate the robustness of different weight combinations in DCCA after adding Gaussian noise of different variances to both the EEG and eye movement features. Figure 7 depicts the results. Although the model achieves the highest classification accuracy when the EEG weight is set to 0.7, it is also more susceptible to noise. The robustness of the model decreases as the weight of the EEG features increases. Since a larger EEG weight leads to more EEG components in the fused features, we might conclude that EEG features are more sensitive to noise than are eye movement features.

Next, we compare the robustness of different models under Gaussian noise with different variances. Taking both classification performance and robustness into consideration, we use DCCA with an EEG weight set to 0.5. Figure 8 shows the performances of the various models. The performance decreases with increasing variances of the Gaussian noise. DCCA obtains the best performance when the noise is lower than or equal to $\mathcal{N}(0,1)$. The performance of the fuzzy integral fusion strategy exceeds DCCA when the noise is stronger than or equal to $\mathcal{N}(0,3)$. The BDAE model performs poorly under noisy conditions even when minimal noise is added to

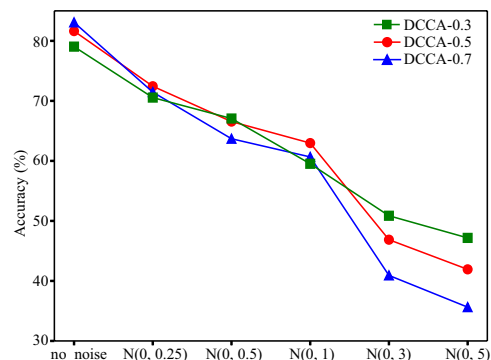the training samples, the performance of the BDAE model is greatly reduced.

Fig. 7. Performance of DCCA with different weight combinations when adding Gaussian noise of different variances. The robustness of DCCA decreases as the weight of the EEG features increases.
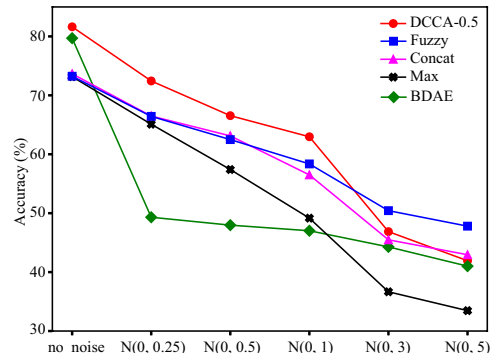

Fig. 8. Model performances after adding Gaussian noise of different variances. The accuracies drops after noise is added to the original training data. DCCA obtains the best performance when the noise is less than $\mathcal{N}(0,1)$. When the noise is stronger than $\mathcal{N}(0,3)$, the fuzzy integral fusion strategy performed best.

2) Replacing EEG features with noise: Table VIII shows the detailed emotion recognition accuracies and standard deviations after replacing 10%, 30%, and 50% percent of the EEG features with different noise distributions. The recognition accuracies decrease with increasing noise proportions. In addition, the performances of seven different settings under different noise distributions are very similar, indicating that

TABLE VIII
RECOGNITION RESULTS (MEAN/STD (%)) AFTER REPLACING DIFFERENT PROPORTIONS OF EEG FEATURES WITH VARIOUS TYPES OF NOISE. FIVE
FUSION STRATEGIES UNDER VARIOUS SETTINGS ARE COMPARED, AND THE BEST RESULTS FOR EACH SETTING ARE IN BOLD

Methods	No noise	Gaussian			Gamma			Uniform		
		10%	30%	50%	10%	30%	50%	10%	30%	50%
Concatenation	73.65/8.90	70.08/8.79	63.13/9.05	58.32/7.51	69.71/8.51	62.93/8.46	57.97/8.14	71.24/10.56	66.46/9.38	61.82/8.35
MAX	73.17/9.27	67.67/8.38	58.29/8.41	51.08/7.00	67.24/10.27	59.18/9.77	50.56/6.82	67.51/9.72	60.14/9.28	52.71/7.84
FuzzyIntegral	73.24/8.72	69.42/8.92	62.98/7.52	57.69/8.70	69.35/8.70	62.64/8.90	57.56/7.19	69.16/8.16	64.86/9.37	60.47/8.32
BDAE	79.70/**4.76**	47.82/7.77	45.89/7.82	44.51/7.43	45.27/**6.68**	45.75/7.91	45.09/8.37	46.13/8.17	46.88/7.14	45.50/9.59
DCCA-0.3	79.04/7.32	76.57/7.63	**73.00**/7.36	**69.56**/7.02	76.87/7.99	**73.06**/7.00	**70.03**/7.17	75.69/**6.34**	**73.22**/6.50	**70.01**/6.66
DCCA-0.5	81.62/6.95	**77.92**/**6.63**	71.77/6.55	65.21/6.24	**78.29**/7.38	72.45/6.14	65.75/6.08	**78.28**/7.16	73.20/6.96	68.01/7.08
DCCA-0.7	**83.08**/7.11	76.27/7.02	68.48/**5.54**	57.63/**5.15**	76.82/7.01	68.54/**6.02**	58.58/**5.44**	77.39/8.43	69.80/**5.63**	61.58/**5.38**

noise distributions have limited influences on the recognition accuracies.

To better observe the changing tendency, we plot the average recognition accuracies under different noise distributions with the same noise ratio. Figure 9 shows the average accuracies for DCCA with different EEG weights. It is obvious that the performances decrease with increasing noise percentages and that the model robustness is inversely proportional to the ratio of the EEG modality. This is the expected performance. Since we only randomly replace EEG features with noise, larger EEG weights will introduce more noises to the fused features, resulting in a decrease in model robustness.

Similar to Fig. 7, we also take DCCA-0.5, as a compromise between performance and robustness to compare with other multimodal fusion methods. Figure 10 depicts the trends of the accuracies of several models. It is obvious that DCCA performs the best, the concatenation fusion achieves a slightly better performance than the fuzzy integral fusion method, and the BDAE model again presents the worst performance.

Combining Figs. 8 and 10, DCCA obtains the best performance under most noisy situations, whereas the BDAE model performed the worst under noisy conditions. This might be caused by the following:

- As already discussed in previous sections, DCCA attemps to preserve emotion-related information and discard irrelevant information. This property prevents the model performance from rapidly deteriorating by neglecting negative information introduced by noise.
- The BDAE model minimizes the mean squared error which is sensitive to outliers [64]. The noisy training features will cause the weights to deviate from the normal range, resulting in a rapid decline in model performance.

D. DREAMER Dataset

For DCCA, we choose the best output dimensions and weight combinations with a grid search. We select the output dimension from the set $[5, 10, 15, 20, 25, 30]$ and the EEG weight α_1 in $[0, 0.1, \cdots, 0.9, 1.0]$ for three binary classification tasks. Figures 11(a), (b), and (c) depict the heat maps of the grid search for arousal, valence, and dominance classifications, respectively. According to Fig. 11, we choose $\alpha_1 = 0.9$ and $\alpha_2 = 0.1$ for the arousal classification, $\alpha_1 = 0.8$ and $\alpha_2 = 0.2$ for the valence classification, and $\alpha_1 = 0.9$ and $\alpha_2 = 0.1$ for the dominance classification.
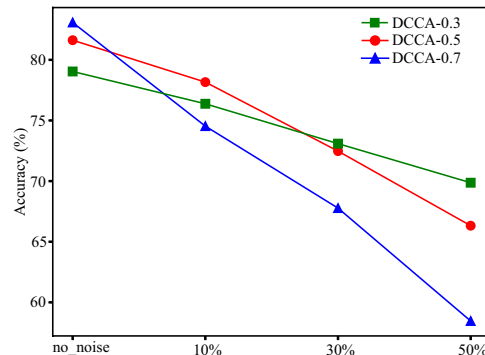

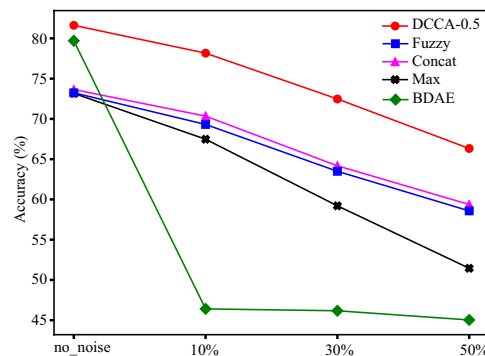Fig. 9. Performance of DCCA with different weight combinations after replacing the EEG features with noise.

Fig. 10. The trends of the average recognition accuracies of different noise distributions under the same noise ratio. The x-axis is the noise replacement ratio, and the y-axis stands for the mean accuracies.

For BDAE, we select the best output dimensions from $[700, 500, 200, 170, 150, 130, 110, 90, 70, 50]$, and leave-one-out cross-validation is used to evaluate the BDAE model.

Table IX gives comparison results of the different methods. Katsigiannis and Ramzan released this dataset, and they achieved accuracy rates of 62.32%, 61.84%, and 61.84% on arousal, valence and dominance classification tasks, respectively [52]. Song and colleagues conducted a series of experiments on this dataset with SVM, graphSLDA, GSCCA, and DGCNN. DGCNN achieved accuracy rates of 85.54% for arousal classification, 86.23% for valence classification, and 85.02% for dominance classification. From Table IX, we can see that BDAE and DCCA adopted in this paper outperform DGCNN. For BDAE, the recognition results for arousal, valence, and dominance are 88.57%, 86.64%, and 89.52%,
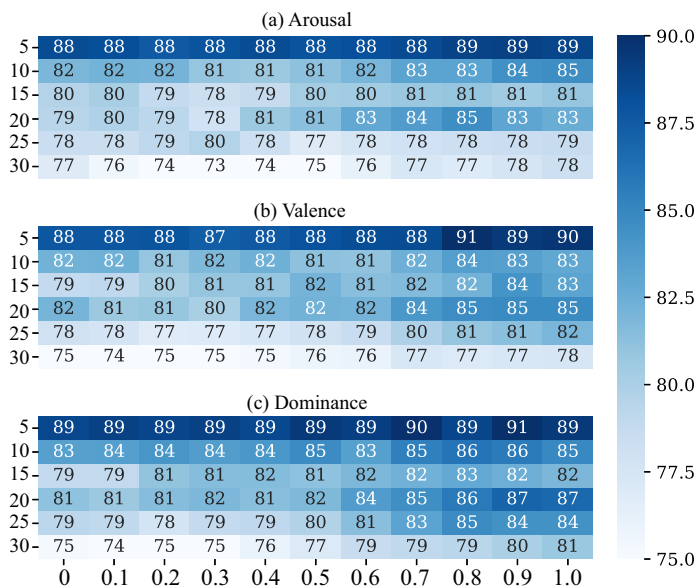
Fig. 11. Selecting the best output dimension and weight combinations of DCCA on the DREAMER dataset. The X-axis represents the weight for the EEG features, and the Y-axis represents the output dimensions.

respectively. DCCA achieves the best performance among all seven methods: 88.99%, 90.57%, and 90.67% for arousal, valence, and dominance level recognitions, respectively.

TABLE IX
COMPARISON OF PERFORMANCES (MEAN/STD, %) ON THE DREAMER DATASET. THREE BINARY CLASSIFICATION TASKS ARE EVALUATED: AROUSAL-LEVEL, VALENCE-LEVEL, AND DOMINANCE-LEVEL CLASSIFICATIONS

Methods	Arousal	Valence	Dominance
Fusion EEG & ECG [52]	62.32/-	61.84/-	61.84/-
SVM [62]	68.84/24.92	60.14/33.34	75.84/20.76
GraphSLDA [62]	68.12/17.53	57.70/13.89	73.90/15.85
GSCCA [62]	70.30/18.66	56.65/21.50	77.31/15.44
DGCNN [62]	84.54/10.18	86.23/12.29	85.02/10.25
BDAE	88.57/4.40	86.64/7.48	89.52/6.18
Our method	**88.99/2.84**	**90.57/4.11**	**90.67/4.33**

VI. CONCLUSION

In this paper, we have introduced deep canonical correlation analysis (DCCA) to multimodal emotion recognition. We have systematically evaluated the performance of DCCA on five multimodal emotion datasets (the SEED, SEED-IV, SEED-V, DEAP and DREAMER datasets) and compared DCCA with the existing emotion recognition methods. Our experimental results demonstrate that DCCA is superior to the existing methods for multimodal emotion recognition.

We have analyzed properties of the transformed features in the coordinated hyperspace $\mathcal{S}$. By applying t-SNE method, we have found qualitatively that: 1) different emotions are better represented since they are disentangled in the coordinated hyperspace; and 2) different modalities have compact distributions from both inter-modality and intra-modality perspectives. We have applied mutual information neural estimation (MINE) algorithm to compare the mutual information of original features and transformed features quantitatively. The experimental results show that the features transformed by DCCA have higher mutual information, indicating that DCCA transformation processes preserve emotion-related information and discard irrelevant information.
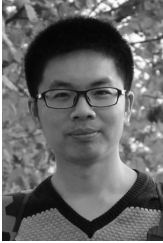
We have investigated the robustness of DCCA on noised datasets under two schemes. By adding Gaussian noise of different variances to both EEG and eye movement features, we have demonstrated that DCCA performs best when the noise is smaller than or equal to $N(0,1)$. After replacing 10%, 30%, and 50% percentage of EEG features with normal distribution, gamma distribution, and uniform distribution, we have shown that DCCA has the best performance for multimodal emotion recognition.

REFERENCES

[1] R. W. Picard, *Affective computing*. MIT press, 2000.
[2] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1175–1191, 2001.
[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
[4] B. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, 2018.
[5] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 25, 2017.
[6] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.
[7] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2017.2712143.
[8] Y. Yang, Q. J. Wu, W.-L. Zheng, and B.-L. Lu, "EEG-based emotion recognition using hierarchical network with subnetwork nodes," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 408–419, 2018.
[9] Z. Yin, Y. Wang, L. Liu, W. Zhang, and J. Zhang, "Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination," *Frontiers in Neurorobotics*, vol. 11, p. 19, 2017.
[10] Y.-P. Lin, J.-H. Chen, J.-R. Duann, C.-T. Lin, and T.-P. Jung, "Generalizations of the subject-independent feature set for music-induced emotion recognition," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 6092–6095.
[11] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from eeg data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.
[12] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 2067–2083, 2008.
[13] M. L.-H. Võ, A. M. Jacobs, L. Kuchinke, M. Hofmann, M. Conrad, A. Schacht, and F. Hutzler, "The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect," *Psychophysiology*, vol. 45, no. 1, pp. 130–140, 2008.
[14] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: from unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.
[15] W.-L. Zheng, W. Liu, Y.-F. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, March 2019.
[16] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.
[17] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, April 2012.

[18] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, Jan 2019.

[19] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, Jan 2016.

[20] Y.-F. Lu, W.-L. Zheng, B.-B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[21] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[22] B. Sun, L. Li, X. Wu, T. Zuo, Y. Chen, G. Zhou, J. He, and X. Zhu, "Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 125–137, 2016.

[23] T. Baltruaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2017.

[24] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 521–529.

[25] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep neural networks," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 811–819.

[26] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2016, pp. 352–359.

[27] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Computer Methods and Programs in Biomedicine*, vol. 140, pp. 93–110, 2017.

[28] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[29] J.-L. Qiu, W. Liu, and B.-L. Lu, "Multi-view emotion recognition using deep canonical correlation analysis," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 221–231.

[30] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[31] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys*, vol. 47, no. 3, pp. 1–36, 2015.

[32] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018, pp. 196–201.

[33] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011, pp. 689–696.

[34] H. Monkaresi, M. Sazzad, and R. A. Calvo, "Classification of affects using head movement, skin color features and physiological signals," in *IEEE International Conference on Systems*, 2012.

[35] K. Guo, R. Chai, H. Candra, Y. Guo, R. Song, H. Nguyen, and S. Su, "A hybrid fuzzy cognitive map/support vector machine approach for EEG-based emotion classification using compressed sensing," *International Journal of Fuzzy Systems*, vol. 21, pp. 263–273, 2019.

[36] I. Naim, Y. C. Song, Q. Liu, H. Kautz, J. Luo, and D. Gildea, "Unsupervised alignment of natural language instructions with video segments," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, 2014, pp. 1558–1564.

[37] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, 2013, pp. 2121–2129.

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

[39] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[40] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[41] A. Klami and S. Kaski, "Probabilistic approach to detecting dependencies between data sets," *Neurocomputing*, vol. 72, no. 1, pp. 39–46, 2008.

[42] A. Klami, S. Virtanen, and S. Kaski, "Bayesian canonical correlation analysis," *Journal of Machine Learning Research*, vol. 14, no. Apr, pp. 965–1003, 2013.

[43] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[44] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Machine Learning*, vol. 83, no. 3, pp. 331–353, 2011.

[45] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Artificial Intelligence and Statistics*, 2014, pp. 823–831.

[46] M. Grabisch and M. Roubens, "Application of the choquet integral in multicriteria decision making," *Fuzzy Measures & Integrals*, pp. 348–374, 2000.

[47] B. Li, X.-C. Lian, and B.-L. Lu, "Gender classification by combining clothing, hair and facial component classifiers," *Neurocomputing*, vol. 76, no. 1, pp. 18–27, 2012.

[48] K. Tanaka and M. Sugeno, "A study on subjective evaluations of printed color images," *International Journal of Approximate Reasoning*, vol. 5, no. 5, pp. 213–222, 1991.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[50] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.

[51] T.-H. Li, W. Liu, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from eeg and eye movement signals: Discrimination ability and stability over time," in *9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 607–610.

[52] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[53] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.

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

[55] L.-C. Shi and B.-L. Lu, "Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 6587–6590.

[56] Y. Hsu, J. Wang, W. Chiang, and C. Hung, "Automatic ecg-based emotion recognition in music listening," *IEEE Transactions on Affective Computing*, pp. 1–16, 2018.

[57] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 2016, pp. 95–108.

[58] A. M. SCHER and A. C. YOUNG, "Frequency analysis of the electrocardiogram," *Circulation Research*, vol. 8, no. 2, pp. 344–346, 1960.

[59] S. A. Shufni and M. Y. Mashor, "Ecg signals classification based on discrete wavelet transform, time domain and frequency domain features," in *2015 2nd International Conference on Biomedical Engineering (ICoBE)*. IEEE, 2015, pp. 1–6.

[60] L. G. Tereshchenko and M. E. Josephson, "Frequency content and characteristics of ventricular conduction," *Journal of Electrocardiology*, vol. 48, no. 6, pp. 933–937, 2015.

[61] L.-M. Zhao, R. Li, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from eeg and eye movement signals: Complementary representation properties," in *9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 611–614.

[62] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, 2018.

[63] J. Chen, B. Hu, Y. Wang, Y. Dai, Y. Yao, and S. Zhao, "A three-stage decision framework for multi-subject emotion recognition using physiological signals," in *IEEE International Conference on Bioinformatics & Biomedicine*, 2017.

[64] J. Kim and C. D. Scott, "Robust kernel density estimation," *Journal of Machine Learning Research*, vol. 13, no. Sep, pp. 2529–2565, 2012.

Bao-Liang Lu (M'94–SM'10) received his B.S. degree in Instrument and Control Engineering from the Qingdao University of Science and Technology, Qingdao, China, in 1982; his M.S. degree in Computer Science and Technology from Northwestern Polytechnical University, Xian, China, in 1989; and his Dr.Eng. degree in Electrical Engineering from Kyoto University, Kyoto, Japan, in 1994.

He was with the Qingdao University of Science and Technology from 1982 to 1986. From 1994 to 1999, he was a Frontier Researcher with the Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research (RIKEN), Nagoya, Japan, and a Research Scientist with the RIKEN Brain Science Institute, Wako, Japan, from 1999 to 2002. Since 2002, he has been a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He received the IEEE Transactions on Autonomous Mental Development Outstanding Paper Award from the IEEE Computational Intelligence Society in 2018. His current research interests include brain-like computing, neural networks, machine learning, brain-computer interaction, and affective computing.

Prof. Lu is currently a Board Member of the Asia Pacific Neural Network Society (APNNS, previously APNNA) and a Steering Committee Member of the IEEE Transactions on Affective Computing. He was the President of the Asia Pacific Neural Network Assembly (APNNA) and the General Chair of the 18th International Conference on Neural Information Processing in 2011. He is currently the Associate Editor of the IEEE Transactions on Cognitive and Development Systems.
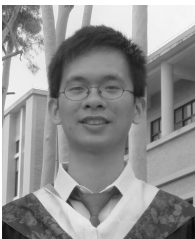
Wei Liu received his bachelor's degree in Automation Science from the School of Advanced Engineering, Beihang University, Beijing, China, in 2014. He is currently pursuing his Ph.D degree in Computer Science from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

His research focuses on affective computing, brain-computer interface, and machine learning.

Jie-Lin Qiu is an undergraduate student at the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China.

His research interests lie in the general area of machine learning, particularly in deep learning and reinforcement learning, as well as their applications in affective computing, brain-machine interfaces, computer vision, and robotics.

Wei-Long Zheng (S'14–M'19) received his bachelor's degree in Information Engineering with the Department of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, in 2012. He received his Ph.D. degree in Computer Science from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2018. Since 2018, he has been a research fellow with the Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. He received the IEEE Transactions on Autonomous Mental Development Outstanding Paper Award from the IEEE Computational Intelligence Society in 2018. His research focuses on affective computing, brain-computer interaction, machine learning and clinical healthcare.