Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition

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

***Abstract*—Multimodal signals are powerful for emotion recognition since they can represent emotions comprehensively. In this paper, we compare the recognition performance and robustness of two multimodal emotion recognition models: deep canonical correlation analysis (DCCA) and bimodal deep autoencoder (BDAE). The contributions of this paper are three folds: 1) We propose two methods for extending the original DCCA model for multimodal fusion: weighted sum fusion and attention-based fusion. 2) We systemically compare the performance of DCCA, BDAE, and traditional approaches on five multimodal datasets. 3) We investigate the robustness of DCCA, BDAE, and traditional approaches on SEED-V and DREAMER datasets under two conditions: adding noises to multimodal features and replacing EEG features with noises. Our experimental results demonstrate that DCCA achieves state-of-the-art recognition results on all five datasets: 94.6% on the SEED dataset, 87.5% on the SEED-IV dataset, 84.3% and 85.6% on the DEAP dataset, 85.3% on the SEED-V dataset, and 89.0%, 90.6%, and 90.7% on the DREAMER dataset. Meanwhile, DCCA has greater robustness when adding various amounts of noises to the SEED-V and DREAMER datasets. By visualizing features before and after DCCA transformation on the SEED-V dataset, we find that the transformed features are more homogeneous and discriminative across emotions.**

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

I. Introduction

EMOTION strongly influences in our daily activities such as interactions between people, decision making, learning, and working. Picard *et al.* 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]. Human emotion recognition is a current hotspot in affective computing research, and it is critical for applications such as affective brain-computer interface [2], emotion regulation and the diagnosis of emotion-related diseases [3].

Traditional emotion recognition systems are built with non-physiological signals [4], [5]. However, emotions also contain reactions from the central and peripheral nervous systems. Besides, electroencephalography (EEG)-based emotion recognition has been demonstrated to be a reliable method because of its high recognition accuracy, objective evaluation and stable neural patterns [6], [7], [8], [9], [10].

In recent years, researchers have tended to study emotions through EEG signals. Various methods have been proposed for EEG-based emotion recognition [11], [12], [13], [14], [15], [16], [17], and one of the reasons is that EEG signals are more accurate and difficult to deliberately change by users. Moreover, other physiological signals such as electromyogram, electrocardiogram, skin conductivity, respiration, and eye movement signals are also used to recognize emotions [18], [19].

Because of the complexity of emotions, it is difficult for single-modality signals to describe emotions comprehensively. Therefore, recognizing emotions with multiple modalities has become a promising method [20], [21], [22], [23]. Many studies indicate that multimodal data can reflect emotional changes from different perspective, which are conducive to building a reliable and accurate emotion recognition model.

Multimodal fusion strategy is one of the key aspects in taking full advantage of multimodal signals. Lu and colleagues employed feature-level concatenation, MAX fusion, SUM fusion, and fuzzy integral fusion to merge EEG and eye movement features [24]. Koelstra and colleagues evaluated the feature-level concatenation of EEG features and peripheral physiological features [25]. Sun *et al.* built a hierarchical classifier by combining both feature-level and decision-level fusion for emotion recognition tasks in the wild [26].

Currently, with the rapid development of deep learning, researchers are applying deep learning models to fuse multimodal signals. Deep-learning-based multimodal representation frameworks can be classified into two categories: multimodal joint representation and multimodal coordinated representation [27]. 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 work of Wei Liu and Bao-Liang Lu was supported in part by the National Natural Science Foundation of China (61976135), the National Key Research and Development Program of China (2017YFB1002501), SJTU Trans-med Awards Research (WF540162605), the Fundamental Research Funds for the Central Universities, the 111 Project, and the China Southern Power Grid (Grant No. GDKJXM20185761). (*Corresponding author: Bao-Liang Lu, bllu@sjtu.edu.cn.)

Wei Liu and Bao-Liang Lu are with the Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering, the Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Brain Science and Technology Research Center, Qing Yuan Research Institute, Shanghai Jiao Tong University, 800 Dongchuan Rd., Shanghai 200240, China, and the Center for Brain-Machine Interface and Neuromodulation, Rui-Jin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin 2nd Rd., Shanghai 200020, China.

Jie-Lin Qiu is with the Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, 15213, USA.

Wei-Long Zheng is with the Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

The multimodal coordinated representation framework learns separate representations for each modality and coordinates them into a hyperspace with constraints between different modalities. Many deep learning models have been applied to emotion recognition in very recent years [28], [29], [30], [31], [32], [33], however, the characteristics these two kinds of models have not yet been fully studied.

In this paper, we compare the recognition performance and robustness of deep canonical correlation analysis (DCCA) [32], [34] and bimodal deep autoencoder (BDAE) [28], [35] for multimodal emotion recognition. DCCA learns separate but coordinated representations for each modality under canonical correlation analysis (CCA) constraints. BDAE, which is a method of multimodal joint representation framework, transforms multiple modalities and jointly learns fused features automatically. The main contributions of this paper on multimodal emotion recognition can be summarized as follows:

1) We propose two multimodal fusion methods to extend the original DCCA model: a weighted-sum fusion and an attention-based fusion. The weighted-sum fusion method allows users to set different weights to different modalities while the attention-based fusion method will calculate the weights adaptively.
2) For the SEED-V dataset, we systemically compare the emotion recognition performance of DCCA with that of BDAE and other existing methods. Then, by visualizing transformed features of DCCA, we find that different emotions are disentangled in the coordinated hyperspace. And finally, we calculate and compare the mutual information of multimodal features before and after DCCA transformation.
3) We compare the robustness of DCCA and BDAE and the conventional multimodal fusion methods on the SEED-V and DREAMER datasets under two conditions: adding noises to multimodal features and replacing EEG features with noises. The experimental results show that DCCA has higher robustness than both the BDAE and traditional methods under most noise conditions.
4) We systematically compare the recognition performance of DCCA and BDAE for multimodal emotion recognition on five benchmark datasets: the SEED, SEED-IV, SEED-V, DEAP, and DREAMER datasets. Our experimental results on these five datasets reveal that both DCCA and BDAE have better performance than traditional multimodal fusion methods for multimodal emotion recognition.

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 of standard DCCA and the proposed weighted-sum fusion and attention-based fusion methods, BDAE, and the baseline models utilized in this paper. The experimental settings are reported in Section IV. Section V presents the experimental comparison results and discussions. Finally, conclusions and future work are given in Section VI.

II. Related Work

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 [36]. 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 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.

A. *Feature-level fusion*

Feature-level fusion is a common and straightforward method to fuse different modalities. The features extracted from various modalities are first combined into a high-dimensional feature and then sent as a whole to the models [24], [25], [35].

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 different classifiers and their combination. Ensemble learning is often used to assemble these classifiers [37]. 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 characteristics of EEG and eye movement features by analyzing confusion matrices [24]. Although rule-based fusion methods are easy to use, the difficulty faced by rule-based fusion is how to design a good rule. 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 [26]. Guo *et al.* built a hybrid classifier by combining fuzzy cognitive map and support vector machine (SVM) to classify emotional states with compressed sensing representation [38].

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 [27].

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 [28], [29] and natural language processing [39].

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 [40]. Andrew and colleagues proposed DCCA method, which is another model under the coordinated representation framework [34].

In recent years, more and more researchers use attention mechanism to fuse multimodal signals [41], [42], [43]. Zhou and colleagues proposed an attention-based bidirectional long-short term memory (LSTM) to deal with relation classification in natural language processing [41]. Zadeh and colleagues applied attention-based fusion in the proposed delta-memory attention network (DMAN) model to handle multi-view sequential learning problems [42]. Li and colleagues proposed multimodal adversarial representation network by combining adversarial learning and attention mechanism for click-through rate prediction problem [43]. In this paper, we propose an attention-based fusion strategy to extend the original DCCA model for emotion recognition.

III. Methods

In this section, we describe the building processes of standard DCCA and the proposed weighted-sum fusion and attention-based fusion methods in Section III-A. The baseline methods used in this paper are introduced in Section III-B.

A. Deep Canonical Correlation Analysis

In this paper, we introduce deep canonical correlation analysis (DCCA) to multimodal emotion recognition. The original DCCA was proposed by Andrew and colleagues [34], and it computes representations of two modalities by passing them through multiple stacked layers of nonlinear transformations. Figure 1 depicts the framework of DCCA used in this paper.

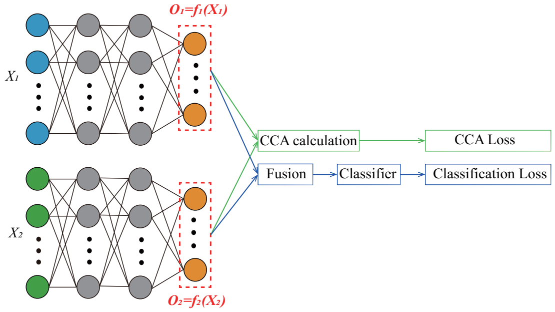


Fig. 1. The framework of the DCCA used in this paper. Different modalities are transformed by different neural networks separately. The outputs (O_1, O_2) are regularized by the traditional CCA constraint. Various strategies can be adopted to fuse O_1 and O_2, and the fused features are used for emotion recognition. We update the parameters to minimize both the CCA loss and the classification loss.

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:

$$O_1 = f_1(X_1; W_1), \tag{1}$$
$$O_2 = f_2(X_2; W_2), \tag{2}$$

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:

$$(W_1^*, W_2^*) = \underset{W_1, W_2}{\arg\max}\ corr(f_1(X_1; W_1), f_2(X_2; W_2)). \tag{3}$$

We use the backpropagation algorithm to update W_1 and W_2. The solution to calculating the gradients of the objective function in Eq. (3) was developed by Andrew and colleagues [34].

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:

$$corr(O_1, O_2) = \left(tr(T'T) \right)^{1/2}. \tag{4}$$

The CCA loss is the negative of total correlation:

$$L_{CCA} = -corr(O_1, O_2) \tag{5}$$

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

$$\frac{\partial corr(O_1, O_2)}{\partial O_1} = \frac{1}{N-1} (2 \nabla_{11} \bar{O}_1 + \nabla_{12} \bar{O}_2), \tag{6}$$

where

$$\nabla_{11} = -\frac{1}{2} \hat{\Sigma}_{11}^{-1/2} U D U' \hat{\Sigma}_{11}^{-1/2}, \tag{7}$$
$$\nabla_{12} = \hat{\Sigma}_{11}^{-1/2} U V' \hat{\Sigma}_{22}^{-1/2}, \tag{8}$$

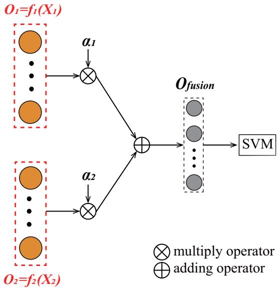


Fig. 2. The process for our proposed weighted sum fusion.

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 [34], 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 extend the original DCCA to fuse multi-modal signals and propose two fusion strategies: 1) weighted sum fusion and 2) attention-based fusion.

1) Weighted sum fusion: For weighted sum fusion, the detailed process for feature fusion and classification is depicted in Figure 2. We initialize two hyper-parameters α_1 and α_2, manually find the best value of these two weights, and fuse different modalities as follows:

$$O = \alpha_1 O_1 + \alpha_2 O_2, \tag{9}$$

where α_1 and α_2 are weights satisfying $\alpha_1 + \alpha_2 = 1$. To find the best combination of weights α_1 and α_2, the grid search method is used to compare the performance of different weight combinations. The α_1 value varies in the range between 0 and 1.0 with a step of 0.1. The grid search results are given in Section V.

And finally, we use SVM to build emotion model with the fused features. Since the tuning of α_1 and α_2 and optimization of SVM can not be optimized with backpropagation, we actually apply a two-stage training process which means that we first optimize the CCA loss and extract transformed features, and then we apply weighted sum fusion and SVM for emotion recognition.

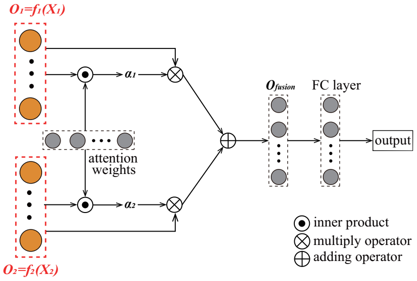


Fig. 3. The process for our proposed attention-based fusion.

2) Attention-based fusion: Figure 3 illustrates the detailed process for our proposed attention-based fusion. First, we initialize an attention layer with parameters W_{attn}, then we calculate the inner product of attention weights and outputs of different modalities and apply softmax to normalize the results getting attention weights α_1 and α_2, respectively.

$$\hat{\alpha}_1 =< O_1, W_{attn} > \tag{10}$$

$$\hat{\alpha}_2 =< O_2, W_{attn} > \tag{11}$$

$$\alpha_1, \alpha_2 = softmax(\hat{\alpha}_1, \hat{\alpha}_2) \tag{12}$$

After calculating the attention weights, we extract the fused features by:

$$O = \alpha_1 O_1 + \alpha_2 O_2 \tag{13}$$

Next, a full-connected (FC) layer is add as classifier with which we can calculate the classification loss. Under attention-based fusion settings, all the updates can be calculated with backpropagation, and we optimize both CCA loss and classification loss simultaneously:

$$L = \gamma_1 L_{CCA} + \gamma_2 L_{classification} \tag{14}$$

where γ_1 and γ_2 are hyper-parameters.

In this paper, we conducted several experiments to discuss the influences of different update ratios $\mathcal{R} = \gamma_1/\gamma_2$. We keep the parameter $\gamma_2 = 1.0$ and choose γ_1 from a set {0.1, 0.3, 0.5, 0.7, 0.9, 1.0} so that the update ratio $\mathcal{R}$ of CCA loss and classification loss ranges from 0.1 to 1.0. We utilize a larger γ_2 since the classification performance is the key metric in the model. Therefore, the penalty of the classification loss should be larger than that of the CCA loss.

According to the construction process mentioned above, the extended DCCA brings the following advantages to multi-modal emotion recognition:

- 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 non-linear mappings ($f_1(\cdot)$ and $f_2(\cdot)$) and make the model preserve the emotion-related information.
- For weighted sum fusion, we assign different priorities to these modalities based on our priori knowledge. In Section V-A, we describe how to find the best α_1 and α_2 and illustrate the influences brought by these two weights.
- For attention-based fusion, we calculate weights for different modalities adaptively. The attention-based fusion can be seen as an adaptive version of the weighted-sum fusion since the weights calculated by attention-based fusion might be the same as weighted-sum fusion and this guarantees that the performance of attention-based fusion will not be worse than that of weighted-sum fusion.

B. Baseline methods

1) Concatenation 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:

$$\begin{aligned} X_{fusion} &= Concat([X^1, X^2]) \\ &= [x_1^1, \cdots, x_n^1, x_1^2, \cdots, x_m^2]. \end{aligned} \tag{15}$$

TABLE I
SUMMARY OF DATASETS AND EXPERIMENTAL SETTINGS.

Dataset	Task	Modality	Training Scheme	Test Scheme
SEED	3 emotions	EEG, Eye movement	session-dependent	train : test=3 : 2
SEED-IV	4 emotions	EEG, Eye movement	session-dependent	train : test=2 : 1
SEED-V	5 emotions	EEG, Eye movement	subject-dependent	3-fold cross-validation
DEAP	2 binary	EEG, peripheral physiological signals	subject-dependent	10-fold cross-validation
DREAMER	3 binary	EEG, ECG	subject-dependent	18-fold cross-validation

2) MAX Fusion: 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:

$$\hat{Y} = \arg\max_{i} \{\arg\max_{j} P_j(Y_i|x_t)\}. \tag{16}$$

3) Fuzzy Integral Fusion: 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

$$\mathcal{C}_\mu(f) := \sum_{i=1}^{n} \big(f(x_{(i)}) - f(x_{(i-1)})\big)\mu(A_{(i)}), \tag{17}$$

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 $f(x_{(0)}) = 0$. We utilize the algorithm proposed by Tanaka and Sugeno [44] to calculate the fuzzy measure.

4) BDAE: BDAE was proposed by Ngiam and colleagues [35]. In our previous work, we adopted BDAE to multimodal emotion recognition [28]. The BDAE training procedure includes encoding and decoding. In the encoding phase, we train two restricted Boltzmann machines (RBMs) for EEG features and eye movement features. 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.

IV. EXPERIMENTAL SETTINGS

In Section 4.1, we introduce the five datasets evaluated in this paper. In Section 4.2, features extraction methods are introduced. And experimental settings are presented in Section 4.3. Table I shows the summary of datasets and experimental settings.

A. Datasets

Five typical multimodal emotion recognition datasets are selected for comparison study in this paper.

1) SEED dataset[1]*:* The SEED dataset was developed by Zheng and Lu [6]. Fifteen Chinese film clips of three emotions (happy, neutral and sad) were used as stimuli in the experiments. Every participant took part in the experiment for three times. In this paper, we use the dataset as in our previous work [24], [28], [29] for the comparison study (9 participants, 27 sessions).The SEED dataset contains EEG signals and eye movement signals.

2) SEED-IV dataset: The SEED-IV dataset was first used in [21]. Seventy-two 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. Fifteen subjects (7 male and 8 female) participated in the experiments for three sessions were performed on different days.

3) SEED-V dataset: The SEED-V dataset was first used in [45]. The dataset contains EEG signals and eye movement signals for five emotions (happy, sad, neutral, fear, and disgust). Sixteen subjects (6 male and 10 female) were required to watch 15 movie clips (3 clips for each emotion), and each of them performed the experiment three times. The SEED-V dataset used in this paper will be freely available to the academic community as a subset of SEED[2].

4) DEAP dataset: The DEAP dataset was developed by Koelstra and colleagues [25]. The EEG signals and peripheral physiological signals of 32 participants were recorded while watching music videos. Participants rated each video on 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 [46]. The DREAMER dataset consists of EEG and electrocardiogram (ECG) signals of 23 subjects (14 males and 9 females). The participants watched 18 film clips to elicit 9 different emotions. After watching a clip, the self-assessment manikins were used to acquire 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 [47], [48].

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 ×

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

TABLE II
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.

$5 = 310$ dimensions for 62 EEG channels. The linear dynamic system method is used to filter out noise and artifacts [49].

For the DEAP dataset, we extract the DE features from four frequency bands: theta, alpha, beta, and gamma (no delta band because the downloaded processed data is filtered to 4-75 Hz.). As a result, there are 128 dimensions for the DE features.

2) ECG feature extraction: In previous work of ECG-based emotion recognition, researchers extracted time-domain features, frequency-domain features, and time-frequency-domain features from ECG signals for emotion recognition [46], [50]. Since there are no standard frequency separation methods for ECG signals [51], we extract the logarithm of the average energy of five frequency bands (1– 4 Hz, 5 – 8 Hz, 9 – 14 Hz, 15 – 31 Hz, and 32 – 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 features extracted from SMI ETG eye-tracking glasses[3] contain both statistical features and computational features. Table II shows all 33 eye movement features used in this paper.

4) Peripheral physiological signal features: For peripheral physiological signals from the DEAP dataset, we calculate statistical features in the temporal domain: the maximum value, minimum value, mean value, standard deviation, variance, and squared sum. For 8 channels of 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 'session-dependent' models for three emotions (happy, sad, and neutral), which is the same as in our previous work [24], [28], [29]. Since every participant took part in the experiment for three sessions, and we build a model for every session, we call the model 'session-dependent' as shown in Table I.

As can be seen from Table I, the test schemes for different datasets are different. The five datasets used in this paper are collected by different research teams at different times. The test schemes for emotion recognition tasks of these datasets are different in the original papers [6], [21], [45], [25], [46]. Most previous studies use the same test schemes as the original papers to report a fair comparison. In this paper, we also use the same test schemes as the original papers to compare our methods with the existing methods.

For SEED-IV dataset, 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 [21]. DCCA is trained under 'session-dependent' setting 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* [52]. We adopt three-fold cross-validation to evaluate the performance of DCCA on a five emotion (happy, sad, fear, neutral, and disgust) recognition task. Since the participant watched 15 movie clips in one session (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) which is a 'subject-dependent' setting.

For the DEAP dataset, we build a subject-dependent model with a 10-fold cross-validation on two binary classification tasks: arousal-level classification and valence-level classification with a threshold of 5.

For the DREAMER dataset, we utilize leave-one-out cross-validation (*i.e.*, 18-fold validation) to evaluate the performance of DCCA, BDAE and baseline methods on three binary classification tasks (arousal, valence, and dominance), which is the same as that used by Song *et al.* [53].

Table III 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. For BDAE model, we use grid search to find the best number of neurons in hidden layers (hidden units are selected from list [200, 150, 100, 90, 70, 50, 30, 20, 15, 10]), and the optimization algorithm is RMSProp with learning rate 0.001. Classifiers for baseline methods mentioned in Section III-B are linear SVM with the same experimental settings as DCCA and BDAE for different datasets.

V. EXPERIMENTAL RESULTS

In this section, we present the experimental results. In Sections V-A and V-B, we examine the effectiveness of DCCA on the SEED-V and DREAMER datasets, respectively. In Section V-C, we compare the recognition performance of DCCA, BDAE, and the traditional multimodal fusion approaches on the SEED, SEED-IV and DEAP datasets. In Sections V-D and V-E, we evaluate the robustness of DCCA, BDAE, and traditional methods on the SEED-V and DREAMER datasets, respectively.

It is worth noting that the weighted-sum fusion method is evaluated on all of the five datasets, while the attention-based

[3] https://en.wikipedia.org/wiki/SensoMotoric_Instruments

TABLE III
SUMMARY OF THE DCCA STRUCTURES FOR FIVE DIFFERENT DATASETS

Datasets	#HiddenLayers	#HiddenUnits	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

method is only evaluated on the SEED-V dataset, and all the analysis and discussion related to DCCA are based on the weighted-sum fusion method. This is because the SEED-V is a newly developed dataset and we want to give a complete comparison on this dataset. Besides, since the attention-based fusion can be seen as an adaptive version of the weighted-sum fusion, the effectiveness of attention-based fusion method can be evaluated on one dataset.

A. *Effectiveness Evaluation of DCCA on the 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).

1) Output dimension and fusion coefficients: 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.* $\alpha_1 = [0, 0.1, 0.2, \cdots, 0.9, 1.0]$ for DCCA. Since $\alpha_1 + \alpha_2 = 1$, we can calculate the weight for the other modality via $\alpha_2 = 1 - \alpha_1$. Figure 4 shows the heat map of the grid search results. Each row 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, and the highest accuracy is marked by a small red circle. According to Figure 4, we set the output dimension to 12 and the weight of the EEG features to 0.7 (*i.e.*, $\alpha_1 = 0.7, \alpha_2 = 0.3$).

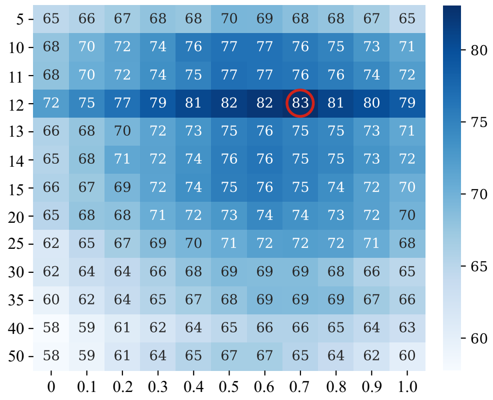


Fig. 4. Selection of the best output dimension and EEG weight of DCCA on the SEED-V dataset. Each row represents the number of output dimensions, and each column denotes the weight (α_1) of the EEG features, and the highest recognition accuracy is marked by a small red circle.

2) Update ratio $\mathcal{R}$ selection: According to experimental settings mentioned in Section III-A2, the update ratio $\mathcal{R}$ ranges from 0.1 to 1.0. Table IV shows the emotion recognition accuracies of SEED-V dataset under different update ratios. From the results, the best performance is obtained with $\mathcal{R} = 0.7$ (i.e. $\gamma_1 = 0.7$ and $\gamma_2 = 1.0$). So in this paper, we set $\gamma_1 = 0.7$ and $\gamma_2 = 1.0$. The setting $\mathcal{R} = 0.7$ can effectively balance the penalty between the CCA loss and the classification loss. Therefore, we use the ratio 0.7 in our further analysis.

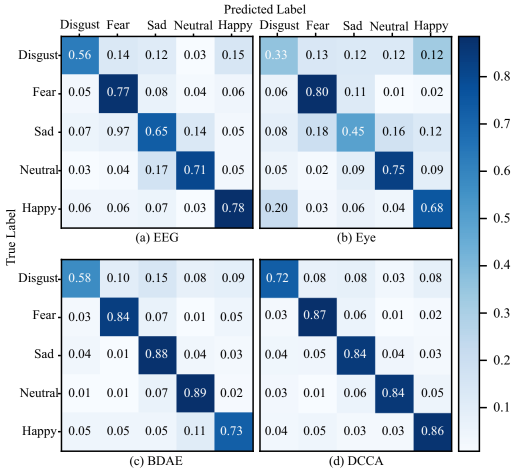


Fig. 5. Comparison of the confusion matrices of different methods on the SEED-V dataset. Subfigures (a), (b), and (c) are the confusion matrices from [52] for SVM classifiers of unimodal features and BDAE model of multimodal features. Subfigure (d) is the confusion matrix of DCCA.

TABLE IV
EMOTION RECOGNITION ACCURACIES UNDER DIFFERENT UPDATE RATIOS

$\mathcal{R}$ (γ_1/γ_2)	0.1	0.3	0.5	0.7	0.9	1.0
Acc (%)	84.5	84.8	84.4	85.3	85.1	84.3
Std (%)	5.5	5.2	4.9	5.6	5.5	5.3

3) Emotion recognition performances: Table V summarizes the emotion recognition results on the SEED-V dataset. Zhao and colleagues [52] adopted feature-level concatenation and BDAE for fusing multiple modalities, and achieved mean accuracy rates of 73.7% and 79.7%, respectively. The MAX fusion and fuzzy integral fusion yielded mean accuracy rates of 73.2% and 73.2%, respectively. The mean accuracy rate of DCCA with weighted-sum fusion is 83.1%, and the result for DCCA with attention-based fusion is 85.3% which is the best result among the six fusion strategies.

Figure 5 depicts the confusion matrices of different methods. Figures 5(a), (b) and (c) are the confusion matrices for the EEG features, eye movement features, and BDAE, respectively. Figure 5(d) depicts the confusion matrix for DCCA. From Figures 5(a), (b), and (d), for each of the five emotions, DCCA achieves a higher accuracy, indicating that

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

Methods	Mean	Std
Concatenation [52]	73.7	8.9
MAX	73.2	9.3
Fuzzy Integral	73.2	8.7
BDAE [52]	79.7	**4.8**
DCCA with weighed-sum fusion	83.1	7.1
DCCA with attention-based fusion	**85.3**	5.6

emotions are better represented and more easily classified in the coordinated hyperspace $\mathcal{S}$ transformed by DCCA.

From Figures 5(a) and (c), compared with the unimodal results of the EEG features, BDAE has worse classification results on the happy emotion, suggesting that BDAE might not take full advantage of different modalities for the happy emotion. Comparing Figures 5(c) and (d), DCCA largely improves the classification results on disgust and happy emotion recognition tasks compared with BDAE, implying that DCCA is more effective in fusing multiple modalities.

4) Visualization of fused features: 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 6 presents a visualization of the features from three participants. Note that the distributions of all participants are similar. Due to the limited space, we only show the distributions of three random subjects. 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 of Figure. 6) 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 more obvious, we separate and plot the distributions of the EEG and eye movement features under the sad emotion in Figure 7. From the perspectives of both inter-modality and intra-modality distributions, the original EEG features ('$\circ$' marker) and eye movement features ('$\times$' 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 $\mathcal{S}$ preserves shared emotion-related information and discards irrelevant information.

Figures 6 and 7 qualitatively indicate that DCCA maps original EEG and eye movement features into a coordinated hyperspace $\mathcal{S}$ where emotions are better represented since only emotion related information is preserved.

5) Mutual information analysis: To support our claims quantitatively, we calculated the mutual information of the original features and transformed features. Figure 8 presents the mutual information of three participants estimated by mutual information neural estimation [54]. The green and red curves depict the mutual information of the original features and the transformed features, respectively. The transformed features have more mutual information than the original features, indicating that the transformed features provide more shared emotion-related information, which is consistent with observations from Figures 6 and 7.

6) Attention weights analysis: As we have mentioned before, attention-based fusion method could calculate weights for EEG features and eye movement features adaptively. Figure 9 shows the average weights of all subjects in SEED-V dataset. The following two observations can be drawn by comparing Figures 9 and 4: 1) EEG features contribute more to the final emotion recognition results than eye movement features and 2) the adaptively computed weights for both EEG features and eye movement features float around the best weights shown in Figure 4, which is consistent with our previous hypothesis that the attention-based fusion could be seen as an adaptive version of weighted-sum fusion.

B. Effectiveness Evaluation of DCCA on the 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 10(a), (b), and (c) depict the heat maps of the grid search for arousal, valence, and dominance classifications, respectively. According to Figure 10, 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.

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 VI gives comparison results of the different methods. Katsigiannis and Ramzan released this dataset, and they achieved accuracy rates of 62.3%, 61.8%, and 61.8% on arousal, valence and dominance classification tasks, respectively [46]. Song and colleagues conducted a series of experiments and compared performance of graph regularized sparse linear discriminant analysis (GraphSLDA), group sparse canonical correlation analysis (GSCCA), and dynamical graph convolutional neural network (DGCNN) on this dataset. The DGCNN method performed better than the other two methods achieving classification accuracy rates of 84.5% for arousal

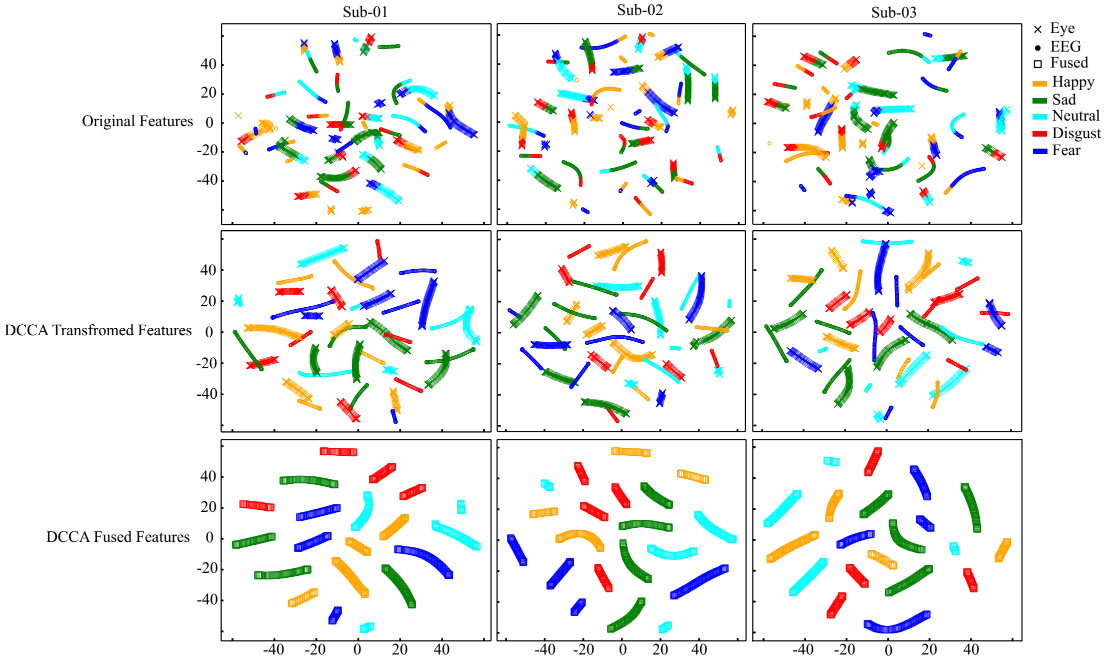


Fig. 6. Feature distribution visualization by the t-SNE algorithm. The original features, transformed features, and fused features from the three subjects are presented. The different colors stand for different emotions, and the different markers indicate different features.

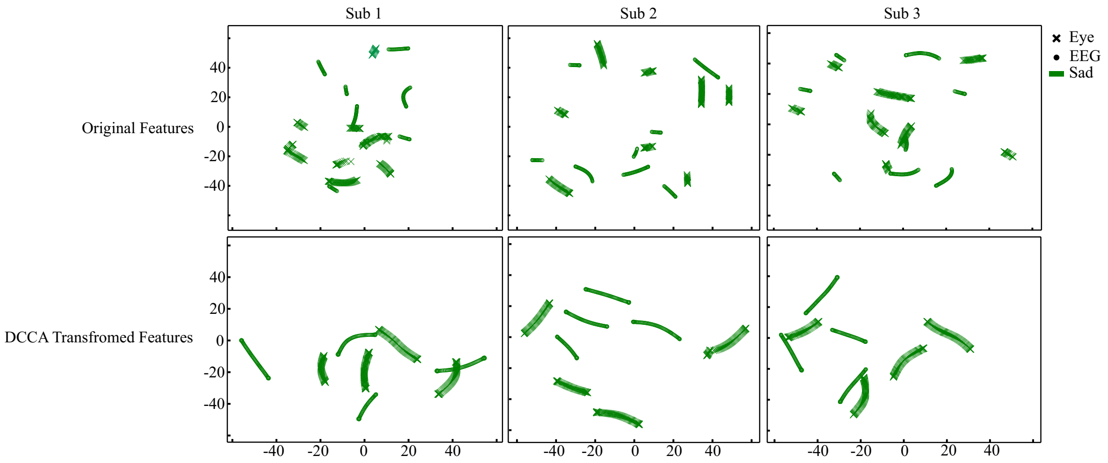


Fig. 7. Distributions of EEG and eye movement features for the sad emotion. The transformed features have more compact distributions from both inter-modality and intra-modality perspectives.

classification, 86.2% for valence classification, and 85.0% for dominance classification [53]. For the concatenation fusion method, the emotion recognition accuracies are 71.4%, 70.1%, and 71.3% for arousal, valence, and dominance classification tasks, respectively. For the MAX fusion method, the emotion recognition accuracies are 72.7%, 72.2%, and 74.3% for arousal, valence, and dominance classification tasks, respectively. The fuzzy integral fusion method achieves 75.7%, 72.4%, and 77.4% accuracies for arousal, valence, and dominance classification tasks, respectively. From Table VI, we can see that BDAE and DCCA adopted in this paper outperform DGCNN. For BDAE, the recognition results for arousal, valence, and dominance are 88.6%, 86.6%, and 89.5%, respectively. DCCA achieves the best performance among all seven methods: 89.0%, 90.6%, and 90.7% for arousal, valence, and dominance level recognitions, respectively.

C. Recognition Performance Comparison

In this section, we present experimental results of DCCA and BDAE on the SEED, SEED-IV, and DEAP datasets. Table VII 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 87.6% [24]. Tang *et al.* [29] adopted bimodal LSTM, obtaining accuracy 94.0%. Recently, Yang and colleagues [8] build a single-layer feedforward network (SLFN) with subnetwork nodes and achieved an accuracy of 91.5%. Song and colleagues [53] proposed DGCNN and obtained a

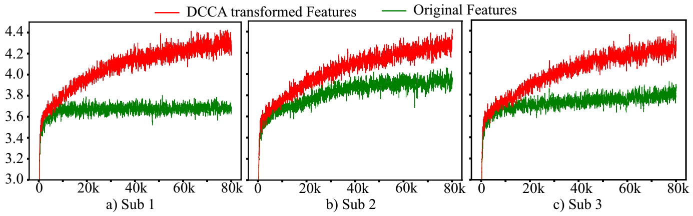


Fig. 8. Mutual information (MI) estimation with MINE. The green curve shows the estimated MI for the original EEG features and eye movement features. The red curve depicts the MI for the transformed features. The x axis is the epoch number of the deep neural network used to estimate MI, and the y axis is the estimated MI. Moving average smoothing is used to smooth the curves.

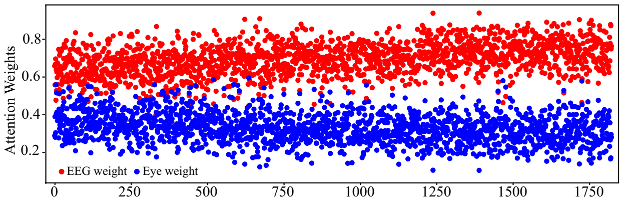


Fig. 9. Visualization of weights calculated by attention-based fusion method. The y-axis shows the calculated weights for different modalities, and the x-axis shows different test samples in the datasets. Red dots stands for EEG weights and blue dots are weights for eye movement features. Similar to Figure 4, EEG features contribute more to the final recognition results and the average weights for all EEG features and eye movement features are also similar to results shown in Figure 4.

classification accuracy of 90.4%. In our previous work [28], BDAE method obtained 91.0% accuracy. From Table VII, we can see that DCCA achieves the best result of 94.6% among the eight different methods.

Table VIII gives the results of five different methods on the SEED-IV dataset. We can observe from Table VIII that for SVM classifier with concatenation fusion, MAX fusion and fuzzy integral fusion, the four emotion states are recognized with a 77.6% mean accuracy rate at the very most. BDAE obtains a mean accuracy rate of 85.1%. DCCA outperforms the aforementioned two methods, with an 87.5% mean accuracy rate.

For DEAP dataset, Table IX shows the results of two binary classifications. As we can observe, DCCA achieves the best results in both arousal classification (84.3%) and valence classification (85.6%) tasks.

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

D. 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 BDAE and the traditional multimodal fusion methods, we designed two experimental schemes on noisy datasets:

- 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

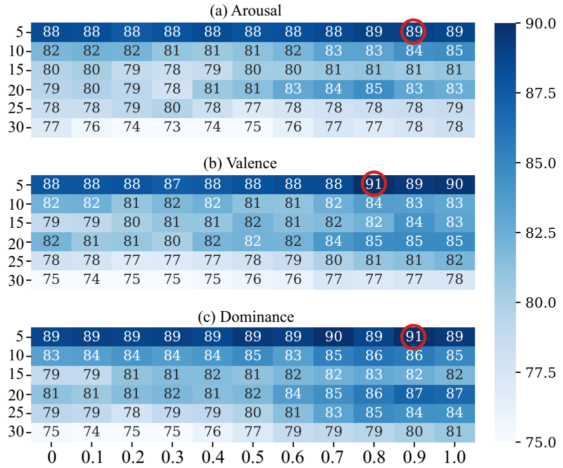


Fig. 10. Selecting the best output dimension and weight combinations of DCCA on the DREAMER dataset. The X-axis represents the weight for the EEG features, and the Y-axis represents the output dimensions. The highest recognition accuracies are marked by a small red circle.

TABLE VI
COMPARISON OF RECOGNITION ACCURACY (MEAN/STD, %) ON THE DREAMER DATASET. THREE BINARY CLASSIFICATION TASKS ARE EVALUATED: AROUSAL-LEVEL, VALENCE-LEVEL, AND DOMINANCE-LEVEL CLASSIFICATIONS. '–' MEANS THE RESULT IS NOT REPORTED.

Methods	Arousal	Valence	Dominance
SVM [46]	62.3/–	62.5/–	61.8/–
SVM [53]	68.8/24.9	60.1/33.3	75.8/20.8
GraphSLDA [53]	68.1/17.5	57.7/13.9	73.9/15.9
GSCCA [53]	70.3/18.7	56.7/21.5	77.3/15.4
DGCNN [53]	84.5/10.2	86.2/12.3	85.0/10.3
Concatenation	71.4/8.2	70.1/10.8	71.3/9.7
Max	72.7/8.4	72.2/7.6	74.3/6.7
Fuzzy Integral	75.7/7.2	72.4/8.9	77.4/6.6
BDAE	88.6/4.4	86.6/7.5	89.5/6.2
DCCA	**89.0/2.8**	**90.6/4.1**	**90.7/4.3**

TABLE VII
THE MEAN ACCURACY RATES (%) AND STANDARD DEVIATIONS (%) OF SEVEN EXISTING METHODS AND DCCA ON THE SEED DATASET. '–' MEANS THE RESULT IS NOT REPORTED.

Methods	Mean	Std
Concatenation [24]	83.7	–
MAX [24]	81.7	–
Fuzzy Integral [24]	87.6	19.9
DGCNN [53]	90.4	8.5
SLFN with subnetwork nodes [8]	91.5	–
Bimodal-LSTM [29]	94.0	7.0
BDAE [28]	91.0	8.9
DCCA	**94.6**	**6.2**

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

Methods	Mean	Std
Concatenation	77.6	16.4
MAX	60.0	17.1
Fuzzy Integral	73.6	16.7
BDAE [21]	85.1	11.8
DCCA	**87.5**	**9.2**

TABLE IX
THE MEAN ACCURACY RATES (%) AND STANDARD DEVIATION (%) OF THREE EXISTING METHODS AND DCCA FOR THE TWO BINARY EMOTION CLASSIFICATION TASKS ON THE DEAP DATASET. '–' MEANS THE RESULT IS NOT REPORTED.

Methods	Arousal	Valence
MESAE [31]	84.2/–	83.0/–
Bimodal-LSTM [29]	83.2/2.6	83.8/5.0
BDAE [28]	80.5/3.4	85.2/4.5
DCCA	**84.3/2.3**	**85.6/3.5**

larger than the generated noise (which is mostly less than 1).

- 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 under a normal distribution ($X \sim \mathcal{N}(0,1)$), gamma distribution ($X \sim \Gamma(1,1)$), and uniform distribution ($X \sim \mathcal{U}[0,1]$).

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). The reason for choosing these three coefficients combination is that we want to examine the effect of different weight coefficients on the robustness of DCCA. The EEG coefficients of 0.3, 0.5 and 0.7 represent settings where EEG features contribute less than, equal to and larger than eye movement features, respectively.

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 11(a) 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 11(b) shows the performance 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 accuracy rates of BDAE greatly reduced even when minimal noise is added.

2) Replacing EEG features with noise: Table X 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 noise distributions have limited influences on the recognition accuracies.

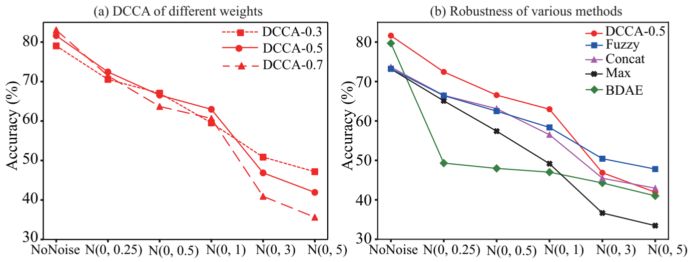


Fig. 11. Performance on the SEED-V dataset of (a) DCCA with different weights and (b) various methods when adding Gaussian noise of different variances.

To better observe the changing tendency, we plot the average recognition accuracies under different noise distributions with the same noise ratio. Figure 12(a) shows the average accuracies for DCCA with different EEG weights. It is obvious that the performance decreases 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 Figure 11(b), we also take DCCA-0.5, as a compromise between performance and robustness to compare with other multimodal fusion methods. Figure 12(b) 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 Figures 11 and 12, DCCA obtains the best performance under most noisy situations, whereas BDAE performs 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.
- BDAE minimizes the mean squared error which is sensitive to outliers [55]. The noisy training features will cause the weights to deviate from the normal range, resulting in a rapid decline in model performance.

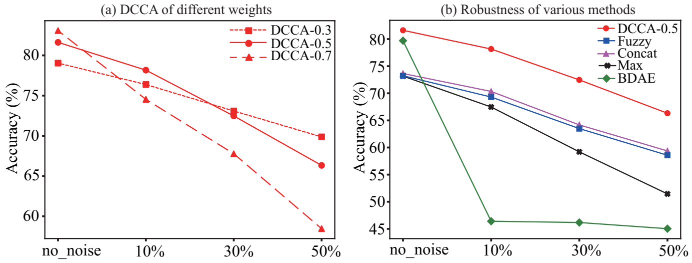


Fig. 12. Performance on the SEED-V dataset of (a) DCCA with different weights and (b) various methods after replacing the EEG features with noise.

TABLE X
RECOGNITION ACCURACY (MEAN/STD (%)) ON THE SEED-V DATASET 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	73.7/8.9	70.1/8.9	63.1/9.1	58.3/7.5	69.7/8.5	62.9/8.5	57.00/8.1	71.2/10.6	66.5/9.4	61.8/8.4
MAX	73.2/9.3	67.7/8.4	58.3/8.4	51.1/7.0	67.2/10.3	59.2/9.8	50.6/6.8	67.5/9.7	60.1/9.3	52.7/7.8
Fuzzy Integral	73.2/8.7	69.4/8.9	63.0/7.5	57.7/8.7	69.4/8.7	62.6/8.9	57.6/7.2	69.2/8.2	64.9/9.4	60.5/8.3
BDAE	79.7/**4.8**	47.8/7.8	45.9/7.8	44.5/7.4	45.3/**6.7**	45.8/7.9	45.1/8.4	46.1/8.2	46.9/7.1	45.5/9.6
DCCA-0.3	79.0/7.3	76.6/7.6	**73.0**/7.4	**69.6**/7.0	76.9/8.0	**73.1**/7.0	**70.0**/7.2	75.7/**6.3**	**73.2**/6.5	**70.0**/6.7
DCCA-0.5	81.6/7.0	**77.9/6.6**	71.8/6.6	65.2/6.2	**78.3**/7.4	72.5/6.1	65.8/6.1	**78.3**/7.2	73.2/7.0	68.0/7.1
DCCA-0.7	**83.1**/7.1	76.3/7.0	68.5/**5.5**	57.6/**5.2**	76.8/7.0	68.5/**6.0**	58.6/**5.4**	77.4/8.4	69.8/**5.6**	61.6/**5.4**

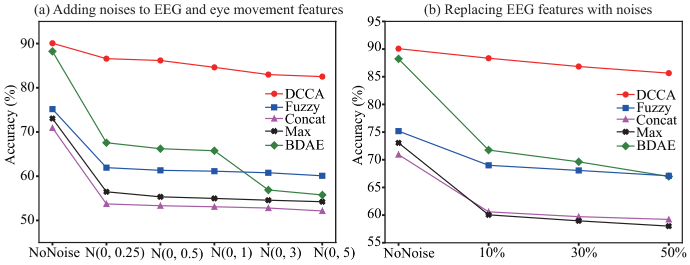


Fig. 13. Model performance on the DREAMER dataset after (a) adding Gaussian noise of different variances and (b) replacing EEG features with noises. The curves shows the average performance of the three binary classification tasks. The x-axis is the type of the Gaussian noise, and the y-axis stands for the mean accuracies.

E. Robustness Analysis on the DREAMER Dataset

In this section, we present the comparison results of robustness of different methods on arousal classification, valence classification, and dominance classification tasks on the DREAMER dataset. Similar to previous settings in Section V-D, we also evaluate the robustness performance under two experimental settings: adding Gaussian noises to both EEG and ECG features and replacing EEG features with noises of Gaussian distribution, gamma distribution, and uniform distribution. For DCCA, we evaluate the robustness performance under the best coefficients combination, *i.e.* $\alpha_1 = 0.9$ for arousal classification, $\alpha_1 = 0.8$ for valence classification and $\alpha_1 = 0.9$ for dominance classification.

1) Adding Gaussian noise: We compare the robustness of different multimodal fusion methods after adding Gaussian noises of different standard deviation $(N(0, 0.25), N(0, 0.5), N(0, 0.1), N(0, 0.3), N(0, 0.5))$ to both EEG and ECG features. Table XI shows the results of arousal, valence and dominance classification tasks after adding various Gaussian noises. From Table XI, we observe that the model performance decreases with the noise standard deviations become larger. In addition, DCCA has better robustness performance than other methods, and BDAE also has a worse performance compared with other methods. The trends of robustness performance of different methods are consistent in all the three tasks.

To better compare the overall performance of different methods, we calculate the average accuracies of all three binary classification tasks under different noise standard deviations. Figure 13(a) shows the average curves of these five multimodal fusion methods. From Figure 13(a), it is obvious that DCCA has the best robustness performance, BDAE has the worst performance, and the concatenation fusion, MAX fusion, and the fuzzy integral methods have similar robustness performance.

Comparing Figure 13(a) and Figure 11(b), the curves in Figure 13(a) change smoother than curves in Figure 11(b) which might be related to the characteristics of different datasets. Since the tasks of the DREAMER dataset are binary classifications, the worst recognition accuracies of noise classifiers tend to be maintained at around 50% leading to a stable change in Figure 13(a).

2) Replacing EEG features with noises: Table XII shows the results of replacing 10%, 30%, and 50% percent of EEG features with Gaussian, Gamma, and Uniform noises for arousal classification, valence classification, and dominance classification. The influences brought by noise types are not very obvious which is consistent with trends shown in Table X.

To better depict the performance of different methods for each of the three binary classification tasks, we first calculate the average performance of the same noise percentage over different noise types and then we calculate the average performance over all three classification tasks. The averaged results as depicted in Figure 13(b). From Figure 13(b), we can see that DCCA performs best since the accuracy reduction is less than other methods, while BDAE has the largest performance gap suggesting a poor robustness. For traditional fusion methods, the fuzzy integral method has better performance than concatenation and MAX fusion methods.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have systematically examined the recognition performance of DCCA, BDAE and traditional methods on five typical multimodal emotion datasets. Particularly, we have proposed two multimodal fusion strategies to extend the original DCCA: a weighted-sum fusion strategy and an attention-based fusion strategy. Our experimental results demonstrate that DCCA is superior to BDAE and the traditional methods for multimodal emotion recognition on all five datasets, and that the attention-based fusion strategy performs better than weighted-sum fusion.

We have analyzed weights from both the weighted-sum fusion strategy and the attention-based fusion strategy. Our experimental results demonstrate that the attention-based fusion strategy can be seen as an adaptive version of the weighted-sum fusion strategy, where the weights calculated by the

TABLE XI
RECOGNITION ACCURACY (MEAN/STD (%)) FOR AROUSAL, VALENCE, AND DOMINANCE CLASSIFICATION TASKS OF THE DREAMER DATASET AFTER ADDING GAUSSIAN NOISES OF DIFFERENT STANDARD DEVIATIONS TO BOTH EEG AND ECG FEATURES.

	Methods	No noise	$N(0,0.25)$	$N(0,0.5)$	$N(0,1)$	$N(0,3)$	$N(0,5)$
Arousal	Concatenation	71.4/8.2	54.0/5.3	53.6/5.8	53.4/7.5	52.7/12.6	52.2/16.9
	MAX	72.7/8.4	60.1/4.8	56.9/8.1	56.2/10.0	55.7/15.1	55.6/17.4
	Fuzzy Integral	75.7/7.2	62.9/5.6	62.6/6.6	62.3/7.6	62.2/8.8	61.9/9.7
	BDAE	88.6/4.4	68.8/7.8	66.8/3.8	67.0/4.6	58.6/15.0	58.6/15.3
	DCCA-0.9	**89.0/2.8**	**87.1/2.8**	**87.0/2.6**	**85.0/2.2**	**83.2/3.3**	**82.8/3.4**
Valence	Concatenation	70.1/10.8	53.7/4.1	53.5/4.8	53.4/6.0	53.3/10.0	52.2/14.5
	MAX	72.2/7.6	54.7/4.1	54.5/5.0	54.4/8.5	54.0/15.1	53.5/18.8
	Fuzzy Integral	72.4/8.9	60.3/4.6	58.9/5.2	58.8/7.1	58.4/11.1	57.6/14.1
	BDAE	86.6/7.5	65.0/8.8	65.6/5.9	64.4/9.2	51.3/22.7	49.9/24.7
	DCCA-0.8	**90.6/4.1**	**85.8/2.9**	**84.8/2.7**	**84.4/3.1**	**83.1/4.7**	**82.2/6.1**
Dominance	Concatenation	71.3/9.7	53.5/3.9	52.9/5.2	52.5/6.6	52.4/10.9	52.0/14.1
	MAX	74.3/6.7	54.7/5.5	54.6/6.3	54.3/8.5	54.0/14.2	53.7/17.1
	Fuzzy Integral	77.4/6.6	62.6/6.0	62.5/7.0	62.3/8.2	61.8/8.3	60.8/9.1
	BDAE	89.5/6.2	68.9/10.4	66.3/5.4	65.9/7.2	60.7/11.6	58.8/11.2
	DCCA-0.9	**90.7/4.3**	**86.9/3.3**	**86.7/3.1**	**84.5/1.8**	**82.7/3.8**	**82.6/3.7**

TABLE XII
RECOGNITION ACCURACY (MEAN/STD (%)) FOR AROUSAL, VALENCE, AND DOMINANCE CLASSIFICATION TASKS OF THE DREAMER DATASET AFTER REPLACING DIFFERENT PROPORTIONS OF EEG FEATURES WITH VARIOUS TYPES OF NOISE.

	Methods	No noise	Gaussian			Gamma			Uniform		
			10%	30%	50%	10%	30%	50%	10%	30%	50%
Arousal	Concatenation	71.4/6.2	61.1/4.8	59.1/5.4	58.7/5.1	61.2/5.9	60.8/5.3	60.7/5.6	60.4/6.0	59.4/5.3	59.0/5.8
	MAX	72.7/8.4	60.6/6.6	59.6/6.9	58.1/6.3	61.3/6.8	59.5/6.9	58.2/5.8	60.7/7.4	60.3/7.1	58.9/7.1
	Fuzzy Integral	75.7/7.2	69.5/4.8	67.9/5.6	66.5/6.3	69.5/5.2	68.3/6.2	66.8/6.7	68.7/5.5	67.3/5.6	67.1/5.8
	BDAE	88.6/4.4	73.8/6.9	73.5/6.4	68.9/7.0	70.6/11.5	70.0/8.3	69.5/7.8	69.5/11.7	68.9/10.0	67.5/9.9
	DCCA-0.9	**89.0/2.8**	**89.1/2.5**	**87.7/2.4**	**85.7/2.8**	**88.6/2.1**	**87.3/2.2**	**86.1/2.6**	**87.6/2.6**	**86.9/2.4**	**85.7/2.2**
Valance	Concatenation	70.1/10.8	60.1/4.5	58.5/4.1	58.1/4.1	60.0/4.9	59.0/4.4	58.5/4.2	60.5/3.8	59.2/4.1	57.2/4.2
	MAX	72.2/7.6	59.5/4.8	58.1/4.6	57.8/5.3	59.2/5.2	57.8/5.1	56.4/5.6	58.3/4.9	57.4/4.5	57.0/5.5
	Fuzzy Integral	72.4/8.9	68.9/4.7	67.8/5.0	66.7/6.3	68.6/4.3	68.3/5.4	67.4/6.8	67.1/4.6	66.9/4.4	66.7/5.4
	BDAE	86.6/7.5	75.1/8.1	70.7/8.7	65.2/8.7	69.9/10.1	68.4/9.9	66.4/10.1	68.6/10.0	67.2/11.0	65.3/10.6
	DCCA-0.8	**90.6/4.1**	**87.2/2.7**	**86.4/2.8**	**85.0/2.7**	**88.3/3.2**	**86.8/3.0**	**85.4/3.0**	**87.8/2.8**	**86.6/3.1**	**85.5/2.9**
Dominance	Concatenation	71.3/9.7	60.7/5.8	60.4/5.2	60.3/5.6	60.5/5.0	60.4/5.5	60.4/5.9	60.8/5.4	60.5/5.0	60.1/4.8
	MAX	74.3/6.7	60.2/5.9	59.5/6.2	58.7/6.3	60.7/6.5	59.6/6.6	58.8/5.9	59.8/5.7	59.0/5.8	58.3/6.7
	Fuzzy Integral	77.4/6.6	69.6/4.8	67.4/4.7	66.9/5.8	69.6/4.8	69.6/4.8	67.0/6.2	69.2/4.6	69.1/4.3	68.7/4.8
	BDAE	89.5/6.2	76.9/8.0	71.1/9.3	67.9/9.0	71.4/9.7	70.1/9.2	67.3/8.7	70.0/10.0	66.9/11.5	65.0/10.3
	DCCA-0.9	**90.7/4.3**	**88.9/3.5**	**86.8/2.2**	**85.7/2.8**	**89.0/8.9**	**86.6/2.5**	**85.9/2.5**	**88.5/4.4**	**86.6/2.9**	**85.9/1.9**

attention-based fusion float around the best weights from the weighted-sum fusion.

We have analyzed properties of the transformed features in the coordinated hyperspace of DCCA. By applying the 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. Our experimental results indicate 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 compared the robustness of DCCA and BDAE on the SEED-V and DREAMER datasets under two schemes: 1) adding Gaussian noise of different variances to both EEG and eye movement features (or ECG features) and 2) replacing 10%, 30%, and 50% percentage of EEG features with different types of noise, the experimental results indicate that DCCA possesses the strongest robustness to noise data among all of the methods.

Although our extensive comparison results indicate that DCCA in recognition performance and robustness is significantly superior to both BDAE and the traditional multimodal fusion methods for multimodal emotion recognition, there is still room for improvement in the following aspects. 1) The CCA metric used in this paper can fuse only two modalities which limits the application of the DCCA method in real life where more than two modalities might be fused simultaneously. We have achieved some preliminary results by extending CCA metric to generalized CCA metric [56]. In the future, we will evaluate the performance and robustness of attention-based deep generalized CCA method to fuse different modalities on more datasets. 2) Only one simple attention mechanism was used in this paper. In the future, we will explore different types of attention mechanisms such as co-attention mechanism [57]. 3) We will investigate multimodal fusion strategies by applying tensor-based fusion [58], [59], [60] and generative adversarial networks [61] in the future.

REFERENCES

[1] R. W. Picard, *Affective Computing*. MIT press, 2000.

[2] M. M. Shanechi, "Brain-machine interfaces from motor to mood," *Nature Neuroscience*, vol. 22, no. 10, pp. 1554–1564, 2019.

[3] A. S. Widge, D. A. Malone Jr, and D. D. Dougherty, "Closing the loop on deep brain stimulation for treatment-resistant depression," *Frontiers in Neuroscience*, vol. 12, p. 175, 2018.

[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*, vol. 10, no. 3, pp. 417–429, 2019.

[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] R. Fourati, B. Ammar, J. Sanchez-Medina, and A. M. Alimi, "Unsupervised learning in reservoir computing for EEG-based emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[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] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[13] X. Wu, W.-L. Zheng, and B.-L. Lu, "Identifying functional brain connectivity patterns for EEG-based emotion recognition," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 235–238.

[14] Z. Guo, X. Wu, J. Liu, L. Yao, and B. Hu, "Altered electroencephalography functional connectivity in depression during the emotional face-word stroop task," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056014, jul 2018.

[15] J. Li, S. Qiu, Y. Shen, C. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3281–3293, 2020.

[16] M. G. Machizawa, G. Lisi, N. Kanayama, R. Mizuochi, K. Makita, T. Sasaoka, and S. Yamawaki, "Quantification of anticipation of excitement with a three-axial model of emotion with EEG," *Journal of Neural Engineering*, vol. 17, no. 3, p. 036011, jun 2020.

[17] Y. Zhang, B. Liu, and X. Gao, "Spatiotemporal dynamics of working memory under the influence of emotions based on EEG," *Journal of Neural Engineering*, vol. 17, no. 2, p. 026039, apr 2020.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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, pp. 1170–1176.

[25] 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.

[26] 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.

[27] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Z. Wang, X. Zhou, W. Wang, and C. Liang, "Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 4, pp. 923–934, 2020.

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

[35] 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.

[36] 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.

[37] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. CRC press, 2012.

[38] 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.

[39] 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.

[40] 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.

[41] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 207–212.

[42] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," *arXiv preprint arXiv:1802.00927*, 2018.

[43] X. Li, C. Wang, J. Tan, X. Zeng, D. Ou, D. Ou, and B. Zheng, "Adversarial multimodal representation learning for click-through rate prediction," in *Proceedings of The Web Conference 2020*, 2020, pp. 827–836.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, pp. 1–10, 2018.

[54] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," *International Conference on Machine Learning*, pp. 531–540, 2018.

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

[56] Y.-T. Lan, W. Liu, and B.-L. Lu, "Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism," *International Joint Conference on Neural Networks 2020*, pp. 1–6, 2020.

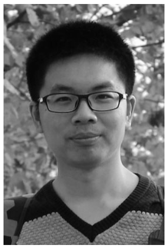
[57] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6087–6096.

[58] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in neural information processing systems*. Citeseer, 2013, pp. 926–934.

[59] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[60] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," *Advances in Neural Information Processing Systems*, vol. 32, pp. 12 136–12 145, 2019.

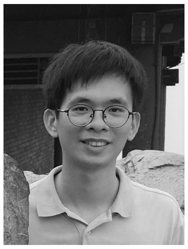
[61] M. Tao, H. Tang, S. Wu, N. Sebe, F. Wu, and X.-Y. Jing, "DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis," *arXiv preprint arXiv:2008.05865*, 2020.



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 received his bachelor's degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2019. He is currently a graduate student in the Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests lie in the general area of machine learning, particularly in deep learning and multimodal machine learning, with their applications in affective computing, brain-machine interfaces, computer vision, robotics, and autonomous systems.



Wei-Long Zheng (S'14-M'19) received the 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 the Ph.D. degree in computer science with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2018. He was a research fellow in the Department of Neurology, Massachusetts General Hospital, Harvard Medical School, USA, in 2018-2020. He is currently a Postdoc Associate in Department of Brain and Cognitive Science at Massachusetts Institute of Technology, USA. He received the IEEE Transactions on Autonomous Mental Development Outstanding Paper Award from IEEE Computational Intelligence Society in 2018. His research focuses on computational neuroscience, affective computing, brain-computer interaction, and machine learning.

Bao-Liang Lu (M'94–SM'01–F'21) received the B.S. degree in instrument and control engineering from Qingdao University of Science and Technology, Qingdao, China, in 1982, the M.S. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 1989, and the 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 network, deep learning, affective brain-computer interface, and emotion artificial intelligence.

Prof. Lu was the President of the Asia Pacific Neural Network Assembly 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 and a Board Member of the Asia Pacific Neural Network Society.