Simplifying Multimodal Emotion Recognition with Single Eye Movement Modality

Xu Yan

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2022

Reading Committee:
Fei Xia, Chair
Qi Cheng

Program Authorized to Offer Degree:
Computational Linguistics

University of Washington

Abstract

Simplifying Multimodal Emotion Recognition with Single Eye Movement Modality

Xu Yan

Chair of the Supervisory Committee:
Professor Fei Xia
Department of Linguistics

Multimodal emotion recognition has long been a popular topic in affective computing since it significantly enhances the performance compared with that of a single modality. Among all, the combination of electroencephalography (EEG) and eye movement signals is one of the most attractive practices due to their complementarity and objectivity. However, the high cost and inconvenience of EEG signal acquisition severely hamper the popularization of multimodal emotion recognition in practical scenarios, while eye movement signals are much easier to acquire. To increase the feasibility and the generalization ability of emotion decoding without compromising the performance, we propose a generative adversarial network-based framework. In our model, a single modality of eye movements is used as input and it is capable of mapping the information onto multimodal features. Experimental results on SEED series datasets with different emotion categories demonstrate that our model with multimodal features generated by the single eye movement modality maintains competitive accuracies compared to those with multimodality input and drastically outperforms those single-modal emotion classifiers. This illustrates that the model has the potential to reduce the dependence on multimodalities without sacrificing performance which makes emotion recognition more applicable and practicable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

COVID-19 makes things harder. I sincerely appreciate everyone who offered me help and guidance during the past three years.

I would like to thank my advisor Fei first. She always listens patiently to my wildest ideas and gives useful insights on what to do next. Without her support, I could not have completed our program successfully.

I also want to thank my reader Qi, who I did research assistant for. Besides the research, she helped me so much in my Ph.D. application and fully backed me to pursue my dream in the field I have passion for.

I really appreciate my advisor, professor Bao-Liang Lu, of my internship during the gap year. Being part of BCMI lab taught me how to be a researcher which mattered a lot when I did the big decision in my life. I'm super grateful for this opportunity.

One of the greatest gains I got from BCMI lab is being mentored by Dr. Liming Zhao. It was him who guided me step by step through the darkness both in research and in life. It is my pleasure to become friends with him.

I appreciate my partner Jessie more than words can say. She is always there when I need with endless patience and love.

Most importantly, I would like to thank my parents. They always trust me and support my decision unconditionally. This gives me great courage to face all difficulties.

Chapter 1

INTRODUCTION

Emotions penetrate our life everywhere and every day and play an important role in the way we think and behave. Accordingly, the emotion intelligence is gradually attracting more attention, especially with the prospects of deep learning. It can be divided into three stages: emotion recognition, emotion understanding, and emotion regulation, offering enormous potential to be used in broad scenarios such as medical diagnosis and treatment, interpersonal relationship improvement, and user experience optimization of general artificial intelligence applications. As the primary step and a salient milestone [5], emotion recognition maintains wide popularity among researchers. Many studies have attempted to find effective modalities to measure emotions, taking facial expression [6, 58, 22], eye movements [45], EEG signals [3, 57], and speech [11, 18] as examples. However, the performances of those individual modalities remain at inadequate levels that cannot be generalized because emotions are complex psychophysiological processes associated with both internal and external activities.

These unsatisfactory findings urge scholars to explore new ways to model the characteristics of emotions. Inspired by the different aspects of information provided by different modalities, some pioneers started to integrate multiple modalities to determine whether the complementary information can help recognize emotions better, and many studies indeed have shown that multimodal fusion emotion recognition methods would achieve better performances than those based on a single modality. Among all groups, the combination of EEG signals, reflecting internal physiological responses, and eye movements, representing external subconscious behaviors, has been proven to be a promising approach with high interpretability [45]. Zheng *et al.* first adopted a multimodal emotion recognition framework by combining these two modalities in three-class emotion recognition (happy, sad and neu-

tral). The impressive experimental results show that the feature-level fusion strategy works well [55]. Liu *et al.* dramatically advanced the state-of-the-art performance of this task by extracting high-level fusion features with a deep neural network model called bimodal deep autoencoder [27].

Although multimodal fusion can achieve better results in emotion recognition, involving more modalities means that there are more possible restrictions in real applications. For example, the process of collecting EEG signals is very complicated. In addition to several inevitable preparations, such as wearing electrode caps and injecting conductive gel, we have to guarantee that the acquisition environment is quiet and without disturbance, since the signals are very subtle and sensitive to interference, thus impeding their use in practical scenarios. Comparatively, other physiological signals are much easier to collect. For instance, eye movement signals can be gathered with small pieces of glasses [28]. Nevertheless, considering the irreplaceability of EEG among all modalities, it is urgent to figure out how to make use of the EEG modality without being constrained by its limitations.

Two solutions stand out from other methods. One is cross-modal transfer using signals from one modality as an input and predicting when being given the other modality [40], and the other is multimodal feature generation based on a single modality. Researchers in the computer vision field innovatively proposed the idea of matching the image features with trained EEG features and therefore use the EEG-based classifier to automatically categorize objects [47]. Jiang *et al.* [20] was the first to apply their ideas in the emotion recognition task. Deep regression neural networks are used to find the regressive connection between the bimodal high-level representations and eye movement features. However, the relationship might not be linear, and this model does not fit the characteristics of the modalities used. Besides, they only examined on one dataset, which is insufficient to judge the performance.

Based on these previous attempts, we decide to map the features from a single modality into high-dimensional multimodal features. In this way, signals from multimodalities are not required as input, while the multimodal knowledge has been encoded into the model in the training stage and assists in emotion recognition. To extract the relationship between eye

movement features and multimodal features, we adopt the bimodal deep autoencoder [34] to get this information. Distinguished from Jiang *et al.* [20], using eye movement features as control conditions, a compact yet effective model originating from the conditional generative adversarial networks (CGANs) can generate the corresponding multimodal features by adversarial learning for each emotion class. We conduct experiments on three SEED series datasets with different numbers of emotion classes (3, 4, and 5) to examine the generalization ability of our model. Extensive experimental results demonstrate that the generative representations achieve competitive performance compared with those using multimodal input, leaving room for reducing modality dependence which makes the technique more practicable.

Chapter 2

RELATED WORK

Multiple physiological modalities have been utilized by researchers to classify emotions. In the literature, there are a surprising number of studies using eye movement signals to perform this task because it not only can observe the users' states naturally and efficiently [4, 45] but also is easy to wear. Lu's comprehensive experimental results prove that pupil diameter, dispersion, fixation duration, saccade duration, saccade amplitude, and nine event statistics are distinguishable for emotions, which could be used as efficient features for emotion recognition [28]. However, even the recognition accuracy provided by the state-of-the-art model is not ideal enough to be used in applications. Meanwhile, another group of researchers focusing on EEG signals were surprised by their potentials in emotion recognition [46, 23]. Alarcao *et al.* [1] conducted a detailed survey about EEG-based emotion recognition, including stimuli, feature extraction, and classifiers.

Considering that emotions are complex psycho-physiological phenomena in nature, scientists turn to build more robust emotion recognition models based on multimodalities that may contain complementary information. This idea has been utilized by researchers from many other fields, including natural language processing (NLP). Text usually teams up with image or audio to boost the performance of models in many tasks. For example, Jain *et al.* [19], Audebert *et al.* [2], and Li *et al.* [25] used semantic information together with visual cues to classify documents. Besides, many studies [32, 14] in captioning and image classification also rely on the cooperation between caption/description and image. Even the traditional NLP task, i.e. machine translation, benefits from multimodal learning by introducing information from other modalities like static images to improve translation quality [7, 52]. For multimodal sentiment analysis, Poria *et al.* [37] and Wang *et al.* [49] proposed

several convolutional neural networks (CNNs)-based approaches while Zadeh *et al.* [53] further improved the performance via an end-to-end fusion method which explicitly represents multimodal interactions between behaviors.

Similar to the group of text and image in NLP, the combination of eye movements and EEG has been gaining much attention in BCIs as a good representative for external behavior activities and internal physiological changes [45, 55, 28]. Zheng *et al.* [55] innovatively combined them and examined on both feature-level fusion and decision-level fusion. The improvement on performances of this groundbreaking work inspired Lu *et al.* [28] to testify their relationships and utilize the advantages of it with various modality fusion strategies in emotion recognition. Besides, Liu *et al.* attempted to use multimodal deep learning techniques to model this task [27]. These studies have suggested that modality fusion seems to be a reliable approach since it significantly enhances the performance compared with a single modality.

Although this complementary collocation achieves satisfying performance, it is unfeasible to put it into large-scale applications due to the inconvenience of data acquisition and inter-subject variability of EEG signals. To overcome this issue, cross-modal transfer learning has a growing body of literature that can be categorized into two threads. One is figuring out the relationship between two modalities, conducting a one-to-one mapping and therefore avoid using one of them in the test stage. The other is to perform joint learning on multimodalities and data generation in training and then use one single modality for testing. Both ways have been supported with successful cases from NLP, computer vision, etc. For example, Scott *et al.* [40] developed a novel deep GANs architecture to effectively bridge the semantic relationship between texts and graphs through transforming visual concepts from characters to pixels. Xu *et al.* [51] also tried to generate images from text based on GANs. DALLE 2 [39] and Imagen [42] are two representatives of the state-of-the-art models on the text-to-image task. Other cross-modal tasks between image and text, e.g. bidirectional generation, also attract many researchers [8, 21].

There have been many projects trying to leverage the rich information from EEG in

multiple research areas. Palazzo *et al.* [35] attempted to generate corresponding images from EEG signals by combining an LSTM recurrent neural network with conditional GANs. At the same time, Spampinato *et al.* [47] conducted an RNN-based method to learn visual stimuli-evoked EEG data and used a CNN-based approach to regress images into the learned EEG representation, thus enabling automated visual classification in a brain-based visual object manifold. Jiang *et al.* [20] pioneeringly tested those methods in the emotion recognition field but in a regressive way, which is not suitable for EEG and leaves room for others to explore more methods. Besides, to the best of our knowledge, there is limited work to address the problems of one-to-more cross-modal transfer in the field of emotion recognition. In the following sections, we will present our proposed method to tackle the fore-mentioned problems.

Chapter 3
METHODS

This chapter walks through each main part of the model and explains structures and the dataflow in details.

3.1 Overview

Our goal is to increase the practicability of the system by utilizing features from a single modality to obtain predictions as accurate as those produced using multimodalities with the help of synthetic multimodal high-dimensional features.

Three categories of features are involved in the whole process, including single-modal features from eye movements, real multimodal features from both EEG and eye movements, and generated multimodal features. As shown in Figure 3.1, the training process can be divided into two phases. Training stage I focuses on multimodal feature fusion to extract real multimodal features with a bimodal deep autoencoder. In training stage II, we adopt conditional generative adversarial networks to generate synthetic multimodal features with single eye movement features. With respect to the test stage, only the eye movement modality is needed as input. The whole process is illustrated in Algorithm 1.

3.2 Training Stage I: Multimodal Feature Extraction

The quality of real multimodal features is of great importance in the whole process. Multimodal feature generation is fundamentally a question of high-dimensional data generation. In general, the features can be engineered by extracting high-dimensional emotion information from eye movement signals and corresponding EEG signals. Specifically, for each sampling point i, we align the eye movement signal $\boldsymbol{X}_{EYE_i} \in \mathbb{R}^n$, where n represents the dimension

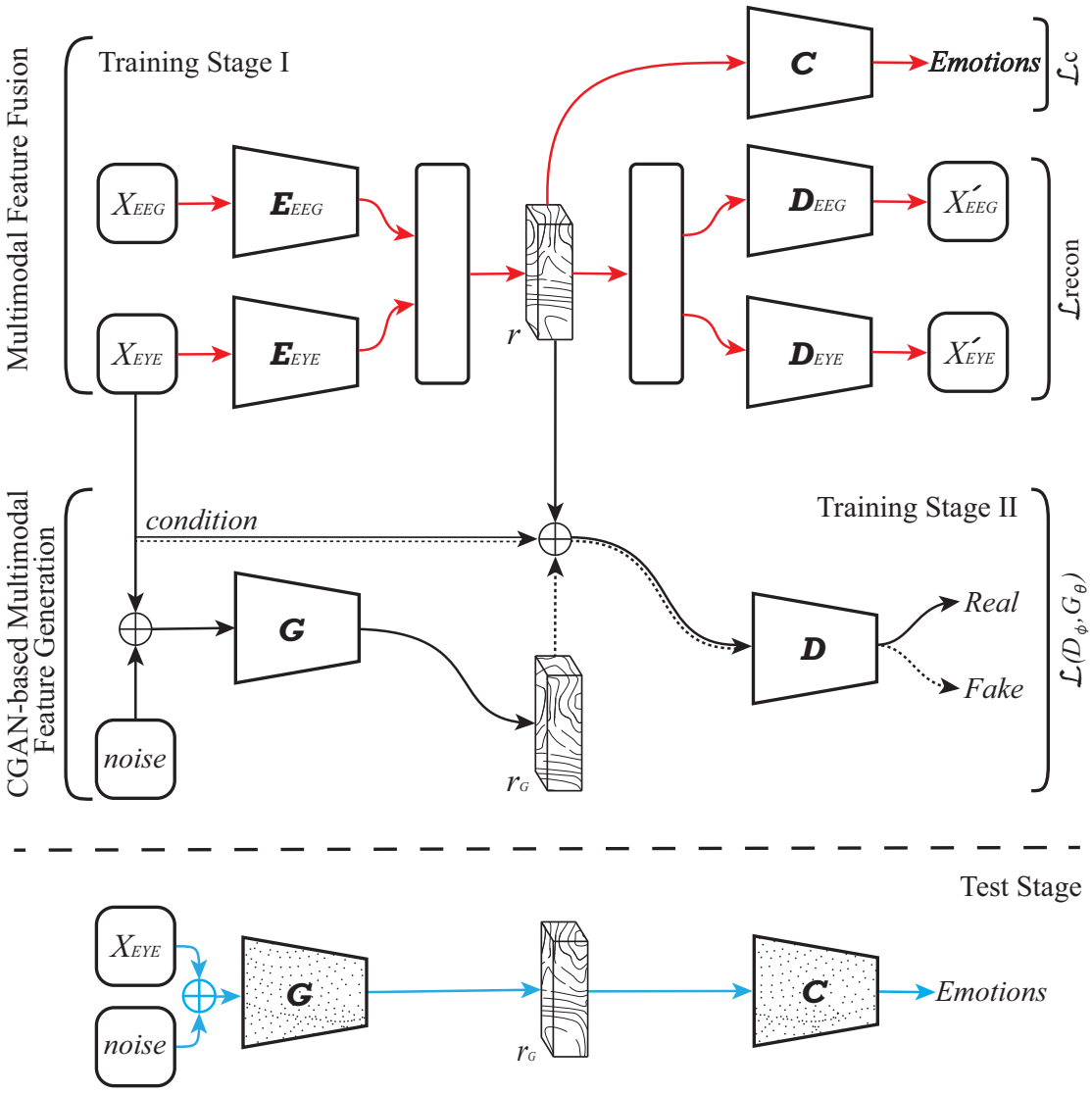

Figure 3.1: The framework of our system. The training stage can be divided into two parts as multimodal feature extraction and multimodal feature generation. In the test phase, only eye movement signals are needed. The shaded $\boldsymbol{G}$ and $\boldsymbol{C}$ indicate that they have been trained.

of the eye movement feature, with the corresponding EEG signal $\boldsymbol{X}_{EEGi} \in \mathbb{R}^m$, where m represents the dimension of the EEG feature.

We choose one of the classic modality fusion methods to extract the high-dimensional

multimodal emotion representations from both EEG and eye movement features, namely the bimodal deep autoencoder [34] shown in training stage I in Figure 3.1 marked with the red lines. There are two steps in the procedure of the bimodal deep autoencoder. The first is encoding, containing two encoders, $\boldsymbol{E}_{EEG}$ for encoding EEG features and $\boldsymbol{E}_{EYE}$ for eye movements. The nature of the encoder is the restricted Boltzmann machine (RBM) [44]. We then train an RBM over the pretrained layers for each modality to model the relationships between them as high-level multimodal representations (r). The second step is decoding. As a process symmetrical to encoding, this step reconstructs the original input representations with $\boldsymbol{D}_{EEG}$ and $\boldsymbol{D}_{EYE}$.

For the emotion classifier, we apply a multilayer perceptron (MLP) as the classifier $\boldsymbol{C}$, which is a feed-forward network. The classifier is trained in the training stage and takes different kinds of features as input. We train the model by minimizing the following loss:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{c}} \tag{1}$$

where the reconstruction loss $\mathcal{L}_{\text{recon}}$ is calculated by the mean squared error:

$$\mathcal{L}_{\text{recon}} = \frac{1}{k} \left\| \boldsymbol{X} - \boldsymbol{X}' \right\|_2^2, \tag{2}$$

where k is the number of features and $\|\cdot\|_2^2$ is the squared L_2-norm. For the cross-entropy loss $\mathcal{L}_c$ of the emotion classifier $\boldsymbol{C}$, the related formula is as follows:

$$\mathcal{L}_{\text{c}} = -\sum_i y_i \log \hat{y}_i \tag{3}$$

where y_i is the ground truth emotion label for input x_i.

Although there are many multimodal encoders to use, our intuition to choose the bimodal deep autoencoder is that the success of the high-dimensional feature decoding process demonstrates that the extracted representations are of high quality and mutually separable in feature space and are qualified to be used to construct a preferable emotion recognition system. Note that in this stage, features from multimodalities are needed.

Algorithm 1: CGAN-based multimodal feature generation algorithm

Input:

EEG data $\boldsymbol{X}_{EEG}$.

Eye movement data $\boldsymbol{X}_{EYE}$.

Divide the training and test sets according to cross-validation.

Output: Recognition accuracy of test data.

Training Stage I:

Initialize encoders $\boldsymbol{E}_{EEG}$ and $\boldsymbol{E}_{EYE}$, decoders $\boldsymbol{D}_{EEG}$ and $\boldsymbol{D}_{EYE}$, and emotion classifier $\boldsymbol{C}$.

for *j=1:n* **do**

    Optimize $\boldsymbol{E}_{EEG}$, $\boldsymbol{E}_{EYE}$, $\boldsymbol{D}_{EEG}$, $\boldsymbol{D}_{EYE}$, and $\boldsymbol{C}$ by minimizing Equation (1).

end

return multimodal features r and trained emotion classifier $\boldsymbol{C}$.

Training Stage II:

Initialize the generator $\boldsymbol{G}$ and discriminator $\boldsymbol{D}$.

for *j=1:n* **do**

    Concatenate random noise and eye movement features in the training set.

    Optimize $\boldsymbol{G}$ and $\boldsymbol{D}$ by minimizing Equation (5).

end

return trained $\boldsymbol{G}$.

Test Phase :

Concatenate random noise and eye movement features in the test set.

Generate the multimodal feature using trained $\boldsymbol{G}$.

Use the trained classifier $\boldsymbol{C}$ for emotion recognition.

return predicted emotion label.

3.3 Training Stage II: Multimodal Feature Generation

Inspired by the broad usage and good performance of generative adversarial networks (GANs) in data augmentation, we select GANs as our basic framework for multimodal feature generation. In the standard GAN structure, a generator $\boldsymbol{G}$ is constructed to produce realistic-like data distribution p_G with mapping function $\boldsymbol{G}(z)$, where z is noise sampled from a noise distribution $p_z(z)$. Besides, a discriminator $\boldsymbol{D}$ is built to play a minimax game against $\boldsymbol{G}$ by distinguishing whether the given sample is in real distribution p_r or generated distribution p_G. The generator-discriminator minimax game for continuous data can be formulated as follows:

$$\min_{G}\max_{D}\mathcal{L}(\boldsymbol{D},\boldsymbol{G}) = \mathbb{E}_{r\sim p_d(r)}[log(\boldsymbol{D}(r))] + \mathbb{E}_{z\sim p_z(z)}[log(1-\boldsymbol{D}(\boldsymbol{G}(z)))] \tag{4}$$

We can see that the standard GAN framework requires the generated data to be differentiable so that the gradient can backpropagate from $\boldsymbol{D}$ to $\boldsymbol{G}$ to update the parameters. This constraint impedes the application of standard GANs to continuous data, i.e., multimodal data. In multimodal data generation tasks, different multimodal data are often associated with the same category, which represents a discrete many-to-one mapping.

This problem can be solved by the conditional probabilistic generative model, where the input is combined with a conditioning variable and generates a conditional predictive distribution. The structure of conditional generative adversarial networks (CGANs) [33] is depicted in Figure 3.1. The eye movement features are taken as conditions to guide the generator to produce the corresponding multimodal features from noise. The discrimination results of $\boldsymbol{D}$ are used to construct the predefined loss function to guide the training of $\boldsymbol{G}$ to guarantee that $\boldsymbol{G}$ can produce realistic and eligible multimodal features. More details of our model are described in the following subsections.

3.3.1 Generator

We employ a fully connected deep neural network (FC-DNN) as our generative model. The FC-DNN generator is composed of multiple fully connected regression layers with *tanh* or

sigmoid activation functions. Other variants of DNN may also be used as the generative model, but a typical FC-DNN can efficiently map the input embedding representation to high-dimensional hidden features, i.e., multimodal emotion features. It is notable that the existing vanishing gradient problem [16] deteriorates the training performance of the generator. This means that during the initial training stage or when $\boldsymbol{D}$ is well learned, $\boldsymbol{D}$ can always reject the generated multimodal features with high confidence so that the gradient guiding $\boldsymbol{G}$ training will approach zero, and the updates of the generator nearly stop. To solve this problem, we improve the FC-DNN model by replacing the *sigmoid* activation function with the $Leaky - ReLU$ activation function for the head layers.

As shown in Figure 3.1, to impose the conditional constraint, we feed the eye movement features $\boldsymbol{X}_{EYE}$ into our FC-DNN generator after concatenation with input noise. The motivation behind this operation is to help $\boldsymbol{G}$ generate multimodal features conditioned based on emotional states. Note that the explicit emotion category is unknown during the real-time emotion classification phase. However, eye movement features capture the emotion information and can be treated as an encoded emotion class label.

3.3.2 Discriminator

The discriminator needs to be a determination function $\boldsymbol{D}(x)$ to generate a single scalar to represent the probability, where x comes from p_r rather than p_G. We also choose FC-DNN as our discriminator. The concatenation of eye movement features constraint $\boldsymbol{X}_{EYE}$ with multimodal features r or r_G as the input is then fed into the discriminator to output the probability that the multimodal feature comes from the real distribution.

Given the generator $\boldsymbol{G}$ and discriminator $\boldsymbol{D}$ with the additional conditional constraint of eye movement features, the objectives of our optimization problem based on the original GAN framework can be rewritten from Equation 4 to Equation 5 as:

$$\begin{aligned}\min_{G_\theta}\max_{D_\phi}\mathcal{L}(\boldsymbol{D}_\phi,\boldsymbol{G}_\theta) =&\mathbb{E}_{(r,\boldsymbol{X}_{EYE})\sim p_d(r)}[log(\boldsymbol{D}_\phi(r,\boldsymbol{X}_{EYE}))]+\\ &\mathbb{E}_{r_G\sim G_\theta(.|\boldsymbol{X}_{EYE})}[log(1-\boldsymbol{D}_\phi(r_G,\boldsymbol{X}_{EYE}))],\end{aligned} \tag{5}$$

where θ and ϕ denote the parameters of the generator and discriminator, respectively, and $\boldsymbol{X}_{EYE}$ is the eye movement constraint. Here, we sample the input noise from the Gaussian noise distribution. The maximum term of $\mathcal{L}(\boldsymbol{D}_\phi, \boldsymbol{G}_\theta)$ (losses of discriminator) and the minimum term of $\mathcal{L}(\boldsymbol{D}_\phi, \boldsymbol{G}_\theta)$ (losses of generator) are optimized in an alternating procedure.

3.4 Test Stage: Emotion Classification

As presented in Figure 3.1, the test data flows according to the blue lines. Using single-modal eye movement features, the trained $\boldsymbol{G}$ can generate corresponding generated multimodal features r_G and pass them to the trained emotion classifier $\boldsymbol{C}$ to predict the emotion.

Chapter 4
EXPERIMENTS

In this chapter, we first introduce the datasets we use. Our implementation details are described in the second session. Finally, we evaluate our model and report the experimental results by comparing with other baselines.

4.1 Public Datasets

To comprehensively verify the performance of our model, we test it on a series of public affective EEG datasets for emotion recognition, including SEED [57], SEED-IV [56], and SEED-V [26] [1]. The basic information of three datasets is listed in Table 4.1.

4.1.1 Stimuli

Several rigorously screened Chinese movie clips are used to elicit the desired target emotion. There are various kinds of stimuli types for emotion elicitation, such as images, texts, music, and videos. Comparatively, videos are informative in both visual and auditory senses and previous studies have validated the reliability and efficiency of films in eliciting emotions [13, 43].

For stimuli generation, according to the dataset publishers, these clips were selected because they were marked with high arousal of the targeted emotion in manual evaluation by another large group of people who did not participate in data collection. Specifically, they picked a collection of movie clips from well-known Chinese movies. Twenty participants were asked to rate their feelings after watching the chosen clips using tags (positive, neutral, and negative) on a scale of 1 to 5. The following criteria were employed to choose the final stimuli

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

Table 4.1: Comparisons of datasets employed. Except for the variables mentioned in this table, other conditions among all three datasets are the same.

Dataset	Subjects	# Emotions	# Videos/exp
SEED[57]	15 (7M8F, 23.27, 2.37)	happy, sad, neutral	15
SEED-IV[56]	15 (7M8F, 23.08, 2.05)	happy, sad, neutral, fear	24
SEED-V[26]	20 (9M11F, 22.15, 1.85)	happy, sad, neutral, fear, disgust	15

from those movie clips: (a) the clips must be easily comprehended without explanation, (b) the duration of the whole experiment shouldn't be too long to make subjects fatigued, and (c) the clips should elicit a single target emotion. In the end, 15 or 24 clips that got a mean rating of 3 or above from the twenty participants were selected from the pool as stimuli [57, 56, 26]. This step is critical to guarantee the intensity of emotion elicitation in formal experiments.

Every clip used in SEED series datasets is 4-minute long and is assigned one emotion label as its gold label based on the content. Gold labels only work for our system and are invisible to subjects during experiments. The clips are equally distributed among different emotion categories.

4.1.2 Protocol

Approximately 15 subjects participated in the experiments three times on different days in order to minimize random errors. During the experiment, subjects are encouraged to immerse themselves in the video to arouse corresponding emotions. The 62-channel EEG signals and the eye movement signals are recorded with the international 10-20 system using the ESI Neuroscan system and SMI′s wearable eye-tracking glasses, respectively, during movie watching. Subjects were instructed to complete a questionnaire right away after watching each clip to report their emotional responses. The questionnaire included the following

questions: (a) how they actually felt after watching the film clip, (b) how they felt at some specific moments, (c) whether they had previously seen the movie, and (d) whether they had understood the clips. Additionally, they were asked to grade the level of their subjective emotional arousal on a five-point scale [57, 56, 26].

4.2 Implementation Details

To compare with other models trained on SEED series datasets, our implementation is consistent with the mainstream in feature extraction and data splitting.

The raw EEG signals are downsampled to 200 Hz and filtered with a bandpass of 0-75 Hz with a baseline correction as well as a PCA-based artifact elimination method with Curry 7 software. Different entropy (DE) features are extracted within a nonoverlapping one-second time window from 5 frequency bands (namely, δ: 1-3 Hz, θ: 4-7 Hz, α: 8-13 Hz, β: 14-30 Hz, and γ: 31-50 Hz) of every sample [10]. Therefore, the dimension of EEG features is 310 per sample, calculated by 62 channels multiplied by 5 bands. We choose DE because various studies have demonstrated that the DE features perform better for EEG-based emotion recognition than other artificial features. For the eye movement signals, 33 features, including pupil diameter, dispersion, fixation duration, blink duration, saccade duration, saccade amplitude, blink frequency, maximum fixation duration and so on, are extracted by SMI BeGaze Analysis software as described in [28, 56].

We use a subject-dependent approach in data splitting, i.e., both training and test data come from the same subject. To improve the generalization ability of the model, we use 3-fold cross-validation and divide the data from one experiment of each subject into three parts on the basis of trials. Two out of three parts are taken as the training set while the remaining is used as the test set each time. Note that considering the time dependency nature of EEG, we did not shuffle the data before splitting.

4.3 Experimental Results

4.3.1 Verification of Feature Generation

Before examining the performance, we start from feature distribution to certify that the structure we construct works properly as we expect. We pick out the generated multimodal features from each subject and visualize them with t-SNE [48], as shown in Figure 4.1. Each

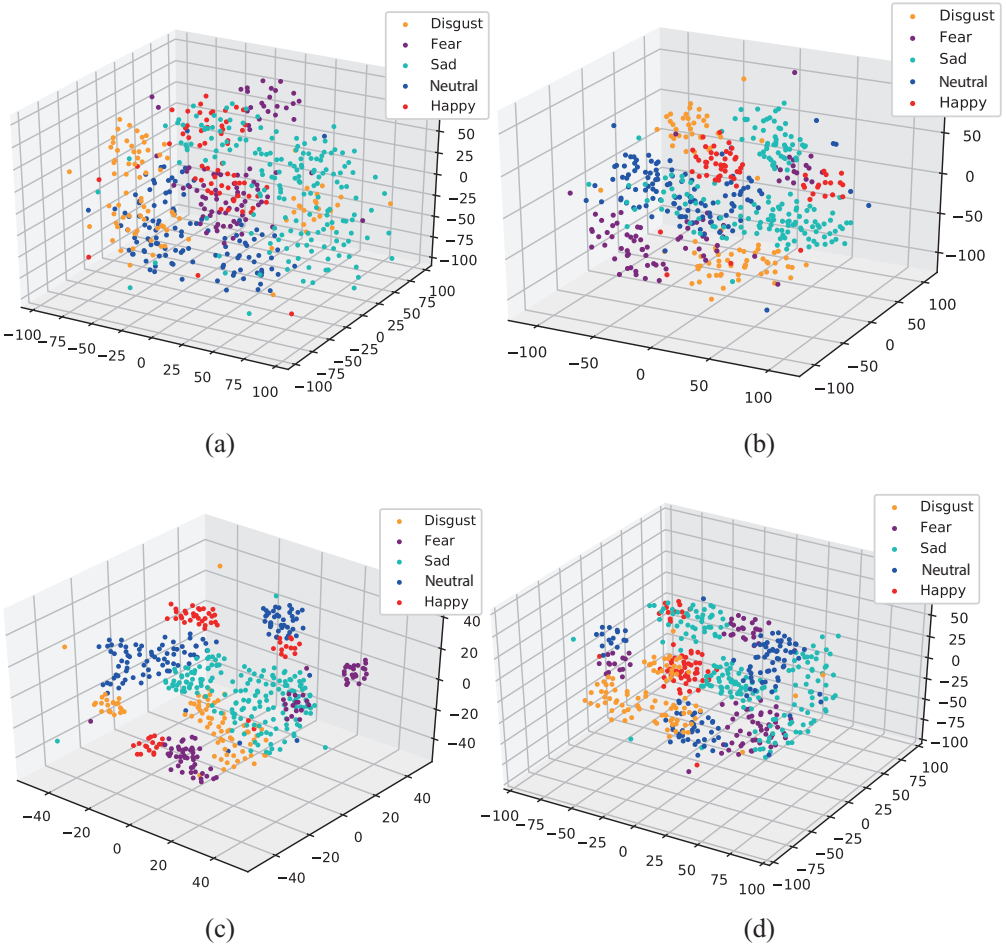

Figure 4.1: Emotion feature visualization, where different colors represent different emotions. (a) Single-modal features of eye movements. (b) Single-modal features of EEG signals. (c) Multimodal features of bimodal deep autoencoder. (d) Generated multimodal features of our model.

point in the figure represents a feature from a 4-second window size. Figure 4.1(a) and Figure 4.1(b) depict features from a single modality as eye movements and EEG, respectively. We can see that there are no obvious clusters in either Figure 4.1(a) or Figure 4.1(b). These results account for the ineffectiveness of emotion classification based on single-modal features. Comparatively, Figure 4.1(c) and Figure 4.1(d), presenting the multimodal features extracted by the bimodal deep autoencoder and those generated by our model, respectively, clearly illustrate the distinguished groups where dots of the same emotion gather. We must admit that the distinction among groups in Figure 4.1(c) is more apparent than in Figure 4.1(d). The visualized distribution further indicates that our method can generate reliable realistic multimodal features, which is a cornerstone to guarantee performance in emotion recognition tasks.

In Table 4.2, the mean accuracies and standard deviations on three datasets of our model are compared with other methods. As demonstrated, the table is separated into three zones. The first two rows are of single modality, while the middle two and the last represent multimodal methods and cross-modal structures, respectively. Besides, we also provide the confusion matrices in Figure 4.2 to directly present the classification results. In general,

Table 4.2: Average accuracies and standard deviations (%) of different methods on 3 datasets.

Feature	SEED		SEED-IV		SEED-V	
Representation	Avg.	Std.	Avg.	Std.	Avg.	Std.
Eye	77.80	14.61	67.82	18.04	59.66	8.77
EEG	78.51	14.32	70.33	14.45	68.58	10.27
Eye+EEG (concatenate)	81.55	11.79	75.88	16.44	73.65	8.90
Eye+EEG (align)	**93.05**	**3.85**	**86.55**	**5.72**	**80.37**	6.03
Eye (regressor)	75.72	8.87	73.49	7.02	72.80	**5.07**
Eye (our work)	81.02	8.04	75.74	6.66	73.66	6.05

the more emotion categories there are, the more difficult the task, which is accompanied by relatively lower performance. Note that our purpose is to build an effective and convenient model with a high generalization ability, so in the detailed discussion below, we devote much attention to the performance on the SEED-V dataset, especially in the figure illustrations.

4.3.2 Comparison with Multimodal Methods

The core issue of multimodal methods is the fusion strategy. Regarding to the ways they fuse, the multimodal methods can be categorized into three types: aggregation-based fusion, alignment-based fusion, and their combination. We select the most commonly used methods among the first two types as representatives to compare with our model. As shown in the third [28] and fourth [27] rows of Table 4.2, alignment fusion performs better since it can effectively extract the relationships among multimodalities. Although our model does not surpass the multimodal alignment fusion method in accuracy, it is as competitive as the one fused by concatenation. Considering that this performance is achieved with only a single modality, we believe our structure can help enhance the user experience and reduce the dependence on modalities. From the confusion matrices in Figure 4.2, we can straightforwardly deduce the superiority of the multimodal method with the darker color. Nevertheless, in the classification of the happy emotion, our model even outperforms the bimodal deep autoencoder by 4% in Figure 4.2(c) and Figure 4.2(d). We also observe that the recognition of disgust is always mixed up with others, especially happy emotion, when the features used are closely related to eye movements as suggested by Figure 4.2(a) and Figure 4.2(d). This might be because eye tracking features such as pupil diameters and saccade details are similar when subjects feel happy and disgusted, which is consistent with the findings of Kuo and Heather [15] that the eye-tracking characteristics of the happy emotion and the disgust emotion have more in common compared with other emotions.

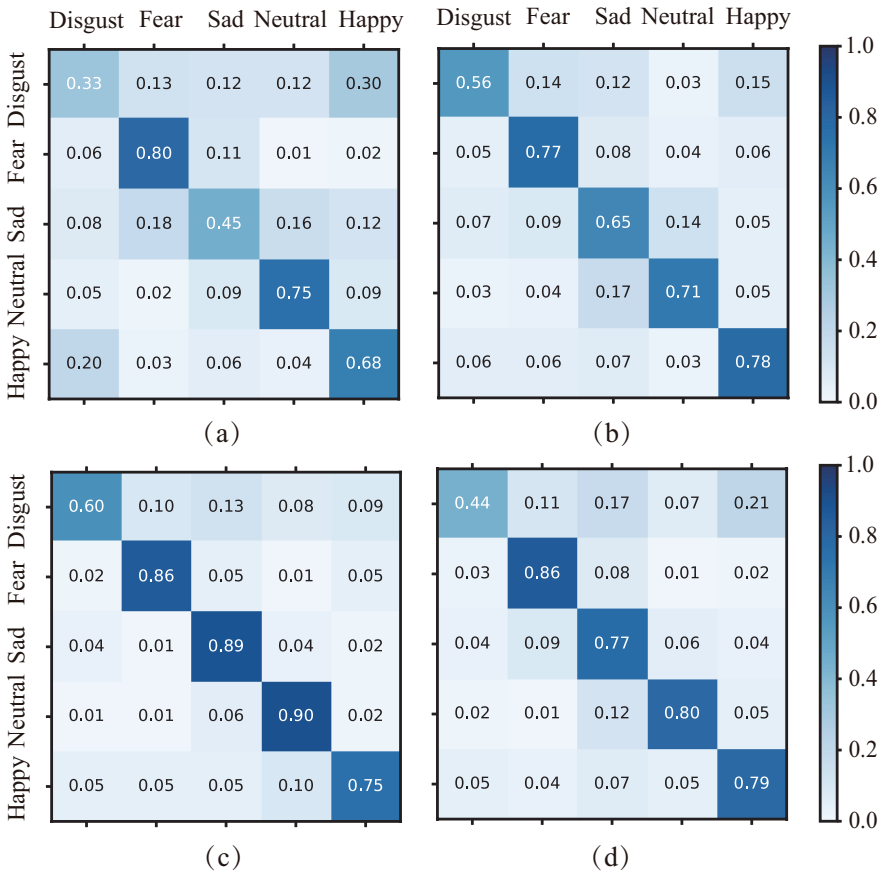

Figure 4.2: Confusion matrices based on the SEED-V dataset. (a) Single-modality of eye movements. (b) Single-modality of EEG. (c) Bimodal deep autoencoder. (d) Our model. The deeper the color, the higher the recognition rate between emotions. The vertical axis represents the true label, and the horizontal axis represents the predicted label.

4.3.3 Comparison with Single-modal Methods

In essence, our system is a single-modal method that only takes the signals from one modality as the input. We compare it with other single-modal models using either eye movements or EEG signals [28] [56], and the results are displayed in the first two rows in Table 4.2. It is apparent from the accuracy values that the EEG signals fit the emotion classification task better and are more reliable than eye movement signals. Regardless of which dataset among the three is employed, our method outperforms all others, with a huge advantage of approximately 5% for EEG signals and 14% for eye movements. This is also reflected in Figure 4.2(a) and Figure 4.2(b), which show that our model outperforms single eye movement in all emotion types. Classification accuracies of the fear and the happy emotion classification have been enhanced by 6% (86% versus 80%) and 11% (79% versus 68%), respectively. Notably, the result of sad emotion recognition using a single eye modality is unsatisfactory. However, with the same input, our model remarkably increases the accuracy by 32%. In addition, the standard deviation is dramatically smaller than those of the existing approaches, indicating that our model is pretty stable and robust. The model we propose has successfully digged out the values of eye movements in emotion recognition tasks, which makes it possible for emotion intelligence to be used in daily life. More importantly, for similar questions in other fields, it provides a new initiative to improve the performance of single-modal methods by integrating knowledge of other modalities.

4.3.4 Comparison with Cross-modal Methods

There are a few cross-modal models used in the emotion recognition task. Compared with the regressor of Jiang [20], it is evident that our model fits the characteristics of data well based on the three relatively higher accuracies. However, although both have a tendency to be more robust when facing more emotion categories, the standard deviations of the regressor decline more quickly than those of our model and even reach 5.07% for the five-class emotion classification task. The relationship between multimodal high-level representations and eye

movement features needs to be further explored in the future.

Chapter 5

CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a new direction to simplify the multimodal structure in emotion recognition tasks by combining a multimodal fusion strategy and a generative model to explore the underlying connections between single modality input and high-dimensional multimodal features. Then, multimodal features can be generated based on a single modality which have the similar characteristics to the real multimodal features extracted from multimodal signals in the training stage. In this way, only information from a single modality is needed in the test stage. The comprehensive experimental results on SEED series datasets demonstrate that our proposed model is as competitive as many multimodal models even though it only requires a single modality as input, which reduces the dependence on multimodalities and makes real-world applications more possible. Moreover, this idea has the potential to be used in other fields to simplify the multimodal process.

For future work, we believe our model can be further improved from the following aspects. First, we plan to embed a data filtering mechanism in the model. This is motivated by the fact that it is hard to guarantee all input data is with high emotion elicitation, even in the most commonly-used public datasets. For example, in SEED dataset, they marked the whole 4-min movie clip as positive. However, people usually need some time to understand what is happening and therefore elicit their emotion, which means that only a part of the clip should be annotated as positive. It will definitely influence the model performance if we input data with inappropriate labels. Zhao *et al.* [54] had investigated this issue by proposing a framework to evaluate the data quality on the basis of spacial-temporal scanpath analysis of eye movements and concluded that data filtering could help enhance the performance. Rather than generating a separate framework, we hope to solve this problem

with a general solution, i.e. teaching the model how to distinguish and filter. Second, we aim to upgrade the model by improving both multimodal fusion and generation parts. Breaking the specification of modalities and expanding the modal coverage is also attractive to us, and the final goal is to build an overall network with which we can use any signal that is the most convenient at the time to perform emotion recognition. Only in this way can we utilize emotion intelligence to benefit people in more practical scenarios such as mental health.

BIBLIOGRAPHY

[1] Soraia M Alarcao and Manuel J Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2017.

[2] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. Multimodal deep networks for text and image-based document classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 427–443. Springer, 2019.

[3] Danny Oude Bos. Eeg-based emotion recognition. *The Influence of Visual and Auditory Stimuli*, 56(3):1–17, 2006.

[4] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008.

[5] Clemens Brunner, Niels Birbaumer, Benjamin Blankertz, Christoph Guger, Andrea Kübler, Donatella Mattia, José del R Millán, Felip Miralles, Anton Nijholt, Eloy Opisso, et al. Bnci horizon 2020: towards a roadmap for the bci community. *Brain-computer Interfaces*, 2(1):1–10, 2015.

[6] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pages 205–211. ACM, 2004.

[7] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *NAACL-HLT (1)*, 2019.

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[10] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering*, pages 81–84. IEEE, 2013.

[11] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[13] James J Gross and Robert W Levenson. Emotion elicitation using films. *Cognition and Emotion*, 9(1):87–108, 1995.

[14] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 902–909. IEEE, 2010.

[15] Kun Guo and Heather Shaw. Face in profile view reduces perceived facial expression intensity: An eye-tracking study. *Acta Psychologica*, 155:19–28, 2015.

[16] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM International Conference on Multimedia*, page 801–804, New York, NY, USA, 2014.

[19] Rajiv Jain and Curtis Wigington. Multimodal document image classification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 71–77. IEEE, 2019.

[20] Huangfei Jiang, Xiya Guan, Wei-Ye Zhao, Li-Ming Zhao, and Bao-Liang Lu. Generating multimodal features for emotion classification from eye movement signals. *Australian Journal of Intelligent Information Processing Systems*, 15(3):59–66, 2019.

[21] Taehoon Kim, Gwangmo Song, Sihaeng Lee, Sangyun Kim, Yewon Seo, Soonyoung Lee, Seung Hwan Kim, Honglak Lee, and Kyunghoon Bae. L-verse: Bidirectional generation between image and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16526–16536, 2022.

[22] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10143–10152, 2019.

[23] Jinpeng Li, Shuang Qiu, Yuan-Yuan Shen, Cheng-Lin Liu, and Huiguang He. Multisource transfer learning for cross-subject eeg emotion recognition. *IEEE Transactions on Cybernetics*, 50(7):3281–3293, 2019.

[24] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.

[25] Pengyuan Li, Xiangying Jiang, Gongbo Zhang, Juan Trelles Trabucco, Daniela Raciti, Cynthia Smith, Martin Ringwald, G Elisabeta Marai, Cecilia Arighi, and Hagit Shatkay. Utilizing image and caption information for biomedical document classification. *Bioinformatics*, 37(Supplement_1):i468–i476, 2021.

[26] Tian-Hao Li, Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Classification of five emotions from eeg and eye movement signals: Discrimination ability and stability over time. In *2019 9th International IEEE/EMBS Conference on Neural Engineering*, pages 607–610. IEEE, 2019.

[27] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal deep learning. In *International Conference on Neural Information Processing*, pages 521–529. Springer, 2016.

[28] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and eeg to enhance emotion recognition. In *International Joint Conference on Artificial Intelligence*, volume 15, pages 1170–1176. Citeseer, 2015.

[29] Yun Luo and Bao-Liang Lu. EEG data augmentation for emotion recognition using a conditional wasserstein GAN. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2535–2538. IEEE, 2018.

[30] Yun Luo, Li-Zhen Zhu, and Bao-Liang Lu. A GAN-based data augmentation method for multimodal emotion recognition. In *International Symposium on Neural Networks*, pages 141–150. Springer, 2019.

[31] Yun Luo, Li-Zhen Zhu, Zi-Yu Wan, and Bao-Liang Lu. Data augmentation for enhancing eeg-based emotion recognition with deep generative models. *Journal of Neural Engineering*, 17(5):056021, 2020.

[32] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *International Conference on Learning Representations*, 2015.

[33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[34] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning*, 2011.

[35] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, and Mubarak Shah. Generative adversarial networks conditioned by brain signals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3410–3418, 2017.

[36] Isola Phillip, Zhu Jun-Yan, Zhou Tinghui, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, Honolulu, HI, USA, 2017.

[37] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544, 2015.

[38] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka, and Kenji Araki. Towards context aware emotional intelligence in machines: computing contextual appropriateness of affective states. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 1469–1474. AAAI, 2009.

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[40] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[41] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 1060–1069, New York City, NY, USA, 2016.

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[43] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7):1153–1172, 2010.

[44] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

[45] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2011.

[46] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

[47] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6809–6817, 2017.

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[49] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2017.

[50] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.

[51] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional

generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[52] Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, 2020.

[53] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[54] Li-Ming Zhao, Xin-Wei Li, Wei-Long Zheng, and Bao-Liang Lu. Active feedback framework with scan-path clustering for deep affective models. In *International Conference on Neural Information Processing*, pages 330–340. Springer, 2018.

[55] Wei-Long Zheng, Bo-Nan Dong, and Bao-Liang Lu. Multimodal emotion recognition using eeg and eye tracking data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5040–5043. IEEE, 2014.

[56] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, (99):1–13, 2018.

[57] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. Identifying stable patterns over time for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, 10(3):417–429, 2017.

[58] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2562–2569. IEEE, 2012.