Multimodal Adaptive Emotion Transformer with Flexible Modality Inputs on A Novel Dataset with Continuous Labels

Wei-Bang Jiang
Shanghai Jiao Tong University
Shanghai, China
935963004@sjtu.edu.cn

Xuan-Hao Liu
Shanghai Jiao Tong University
Shanghai, China
haogram_sjtu@sjtu.edu.cn

Wei-Long Zheng
Shanghai Jiao Tong University
Shanghai, China
weilong@sjtu.edu.cn

Bao-Liang Lu*
Shanghai Jiao Tong University
Shanghai, China
bllu@sjtu.edu.cn

ABSTRACT

Emotion recognition from physiological signals is a topic of widespread interest, and researchers continue to develop novel techniques for perceiving emotions. However, the emergence of deep learning has highlighted the need for high-quality emotional datasets to accurately decode human emotions. In this study, we present a novel multimodal emotion dataset that incorporates electroencephalography (EEG) and eye movement signals to systematically explore human emotions. Seven basic emotions (happy, sad, fear, disgust, surprise, anger, and neutral) are elicited by a large number of 80 videos and fully investigated with continuous labels that indicate the intensity of the corresponding emotions. Additionally, we propose a novel Multimodal Adaptive Emotion Transformer (MAET), that can flexibly process both unimodal and multimodal inputs. Adversarial training is utilized in MAET to mitigate subject discrepancy, which enhances domain generalization. Our extensive experiments, encompassing both subject-dependent and cross-subject conditions, demonstrate MAET's superior performance in handling various inputs. The filtering of data for high emotional evocation using continuous labels proved to be effective in the experiments. Furthermore, the complementary properties between EEG and eye movements are observed. Our code is available at https://github.com/935963004/MAET.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → *Artificial intelligence*; Cognitive science.

KEYWORDS

emotion recognition, continuous label, EEG, eye movements, dataset

*Corresponding author

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613797

ACM Reference Format:
Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. 2023. Multimodal Adaptive Emotion Transformer with Flexible Modality Inputs on A Novel Dataset with Continuous Labels. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3613797

1 INTRODUCTION

Emotion recognition is a critical component of human-computer interactions (HCIs) [16], enabling machines to attain emotional intelligence [41], and allowing computers to identify, understand, and respond to the emotion of human beings. Given the complexity and importance of emotion, a psycho-physiological process triggered by various factors [41], researchers from psychology, neuroscience, and computer science have been exploring emotion recognition for years [2, 36]. However, the challenges of detecting and analyzing human emotions remain largely unexplored.

In recent years, a variety of physiological and non-physiological signals have been utilized for emotion recognition [2]. For non-physiological signals, speech [1, 14, 28], facial expressions [7, 12, 37, 38], and body movements [45, 47], have been utilized by researchers to recognize human emotions. However, non-physiological signals can be easily falsified and are thus untrustworthy, as individuals may conceal their true emotions. In contrast, physiological signals, such as EEG [2, 27, 62], electromyogram (EMG) [23, 39], and electrocardiogram (ECG) [21, 27], provide more reliable and stable options than non-physiological signals. Among all, EEG has shown excellent performance in emotion recognition [2, 62], as it is intrinsically linked to the fundamental neural mechanisms and has been extensively studied in fields such as psychology and neuroscience [10, 58]. Additionally, eye movement signals have been shown to process complementary properties with EEG in multimodal emotion recognition [50, 61]. Therefore, we collect EEG and eye movement signals to create a multimodal dataset.

There are two main approaches to characterizing emotions: the dimensional model and the discrete model. The most well-known dimensional model is the 2D spacial Russell model, where all affective concepts are located at a point with valence and arousal dimensions [43]. Valence represents whether the emotions are positive or negative, while arousal depicts the level of activation or energy associated with an emotional experience. Many emotion

recognition studies have been conducted based on the Valence-Arousal model, such as DEAP [30] and MAHNOB-HCI [50]. Unlike the dimensional approach which portrays emotions continuously, the discrete categorical model, first proposed by Ekman, classifies emotions into a set of discrete statuses [13]. Ekman's theory identified six basic emotions, namely happiness, sadness, fear, disgust, surprise, and anger, which collectively form the basis of all emotional states. The discrete model has also been widely employed in studies such as SEED [62] and research using functional Magnetic Resonance Imaging (fMRI) [44]. Our dataset is based on the discrete model and examines the EEG and eye movement signals of seven emotions, including the six basic emotions and a neutral state.

Compared to intracranial EEG and fMRI, EEG signals are a convenient and non-invasive method for emotion recognition due to their harmlessness, inexpensiveness, and quick acquisition [2]. However, existing EEG emotion datasets such as MAHNOB-HCI [50], DEAP [30], and SEED [62], have limited diversity in emotion types and short recording durations, which restricts their potential for data analysis and performance improvement in emotion recognition. Furthermore, studies on neuroscience and cognitive science have shown that emotions are complex and dynamic physiological processes that vary in intensity and states over time [36]. Therefore, recording continuous intensity labels is a practical way to study these changes. Additionally, multimodal signals have been proven to be effective in emotion classification [33], highlighting the need to record other physiological or non-physiological signals during experiments. To address these challenges, we develop a novel multimodal dataset with continuous labels for emotion recognition focusing on the six basic emotions and the neutral emotion. Our dataset features more than 14,000 seconds of recording time, which is longer than most previous EEG datasets that typically record less than 4,000 seconds.

To address the challenges of emotion classification, we propose a novel Multimodal Adaptive Emotion Transformer (MAET), possessing specialized modules that make it possible for MAET to operate flexibly on both unimodal and multimodal inputs. MAET is first trained with EEG and eye movement features, aiming to learn how to tackle multimodal inputs. Whereafter, we leverage the emotional prompt tuning to enable MAET to recognize emotions using a single modality, while still maintaining the ability to process multimodal features. Moreover, the subject discrepancy is obscured by MAET using adversarial training for promoting domain generalization.

In summary, the main contributions of this paper are as follows:

1) We introduce a novel multimodal emotion dataset focusing on seven basic emotions (happy, sad, fear, disgust, surprise, anger, and neutral), with EEG and eye movement signals recorded. Additionally, continuous labels representing the intensity of the corresponding emotions are collected.
2) We propose a novel Multimodal Adaptive Emotion Transformer (MAET), a flexible model that can process both unimodal and multimodal inputs with specialized modules. Furthermore, Our MAET model alleviates subject discrepancy by adopting adversarial training, thus improving domain generalization.
3) We conduct systematic experiments in various conditions to evaluate MAET compared to other classifiers, including unimodal and multimodal conditions, along with subject-dependent and cross-subject conditions.
4) We analyze the effectiveness of filtering high-induced data using continuous labels. Experimental results indicate that filtering high-induced data can significantly enhance emotion discrimination ability.

2 RELATED WORK

2.1 EEG Dataset for Emotion Recognition

Given the extensive attention that emotion recognition using EEG signals has received, an increasing number of emotional state evaluation methods have been proposed [2]. Hence, comprehensive and high-quality emotional datasets are urgently acquired for researchers to evaluate the performance of their methods. To date, there are several datasets available for classifying emotions that include recordings of EEG along or EEG along with other modalities.

For the Valence-Arousal model, DEAP [30] and MAHNOB-HCI [50] recorded EEG as well as other physiological signals, such as GSR, ECG, and EMG, for emotion research. DEAP revealed that EEG was better for predicting arousal while peripheral physiological signals were better for predicting valence. It is worth mentioning that eye gaze data was proved to be the best single modality for classifying both arousal and valence based on MAHNOB-HCI [50], highlighting the potential effectiveness of eye movement signals in emotion recognition. To make affective computing more applicable in everyday scenarios, wearable and wireless equipment was employed to collect EEG and ECG signals while subjects watched 18 film clips intended to elicit 9 target emotions in DREAMER [27].

Different from the datasets mentioned above, SEED [62] utilized a discrete model to observe the EEG and eye movement states of particular emotions. The dataset selected 15 film clips to evoke positive, neutral, and negative emotions. To acquire high-resolution EEG (HR-EEG) signals, Becker *et. al.* selected 13 videos that consist of 7 positive emotions and 6 negative emotions from FilmStim to obtain HR-EEG along with other physiological signals of 27 subjects. Multiple physiological signals were recorded with 28 emotional videos as elicitation from 23 subjects. Hu *et. al.* constructed THU-EP [22], collecting EEG signals from 80 subjects who responded to 28 video clips of nine emotions, including four positive emotions like joy and amusement, and four negative emotions like anger and disgust. However, existing datasets have some limitations: 1) limited types of emotion states being studied; 2) inadequate videos for inducing each emotion state; and 3) short total video time, which make it difficult to obtain comprehensive and high-quality datasets.

2.2 EEG-based Emotion Recognition

As EEG has been proven to be the most promising physiological signal in emotion recognition, many emotion recognition algorithms based on EEG have been proposed over the years [2]. Zheng *et. al.* employed a deep belief network (DBN) to investigate critical frequency bands and channels of EEG signals for emotion recognition [62]. By reshaping and flattening the EEG signals to image-like tensors according to the spatial relationships, Li *et. al.* used a hierarchical convolutional neural network (HCNN) to learn the spatial pattern of each emotion [31]. Alhagry *et. al.* proposed an EEG feature extraction algorithm using long short-term memory

(LSTM) and applied the features for classifying low/high valence and arousal. To better extract topographical information in EEG signals, a regularized graph neural network (RGNN) which can capture both local and global inter-channel relations was used by Zhong *et. al.* for emotion detection [63]. Song *et. al.* adopted a dynamical graph convolutional neural network (DGCNN) for emotion discrimination which can dynamically learn the intrinsic relationship between EEG channels [52]. Jiang *et. al.* proposed a graph convolutional network with channel attention (GCNCA) to classify anger and surprise emotion [26]. Recently, Transformer has been used for emotion recognition. For example, Wang *et. al.* proposed a Transformer-based model to hierarchically learn the discriminative spatial information [57]. Using attention mechanism on raw EEG signals, Arjun *et. al.* achieved excellent accuracy of 99.4% and 99.1% for classifying valence and arousal [3]. Rajpoot *et. al.* improved LSTM and CNN using attention mechanism for subject-independent emotion recognition [42].

2.3 Multimodal Emotion Recognition

Emotion is an internal subjective experience and is accompanied by various complex imperceptible physiological performances besides facial expressions, such as activation in particular cerebral cortex areas [44] and fluctuation of pupil diameter [60]. Hence, applying multimodal signals can improve discrimination ability and this approach has been widely used in emotion recognition due to the potential complementary properties of different modalities [51, 61].

Sun *et. al.* used a hierarchical classifier with hybrid fusion to distinguish emotions [53]. Fuzzy cognitive map and SVM were employed to form a hybrid classifier for emotion recognition by Guo *et. al.* [19]. A Two-stream heterogeneous graph recurrent neural network was developed to classify emotions, which can fuse spatial-spectral-temporal domain features in a unified framework [25]. With the invention of the attention mechanism, more and more deep methods of fusion have been developed based on the attention mechanism. Liu *et. al.* proposed a deep canonical correlation analysis (DCCA) with an attention-based fusion strategy to perform multimodal emotion recognition [33]. By pre-training Transformers using masked value prediction, Vazquez *et. al.* fused EEG and ECG signals to classify emotions [55]. Nonetheless, these techniques are tailored explicitly for multimodal inputs, and their major drawback is the limited adaptability to unimodal signals.

3 EXPERIMENT SETUP

3.1 Stimuli

The experiments are designed to record EEG and eye movement signals simultaneously during the elicitation of seven emotions. The selection of stimuli materials is critical as it directly affects the effectiveness of emotional elicitation. Previous studies have employed various types of stimuli to evoke emotions, including music [29], pictures [5], facial expressions [7], and videos [30, 50]. Among them, videos have been found to be particularly effective as they provide both visual and auditory stimuli.

During the preliminary stage, a pool of stimuli materials comprising video clips is prepared for eliciting six emotions, excluding surprise. To select the most effective video clips for eliciting emotions, we employ a strategy involving the assessment of all video

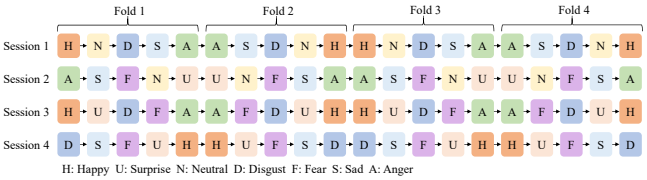


Figure 1: The experiment design of watching videos.

clips by 20 volunteers, who rated each clip on a scale of 1 to 5. We select the high-scoring clips for each emotion. For the emotion of surprise, magic videos are chosen for emotion elicitation as magic shows have been demonstrated to be effective for eliciting surprise [26]. Consequently, twelve clips were selected for each emotion (except neutral) with mean scores of 3 or higher. The neutral emotion comprised eight clips, resulting in a total of 80 clips. Each clip lasts for 2 to 5 minutes and the total time of all clips is about 14,097.86 seconds. We elaborately separate the 80 clips into four sessions, and the subjects are required to complete the entire experiment in four sessions with an interval of 24 hours or more between sessions.

3.2 Subjects

Twenty subjects (10 males and 10 females) with ages ranging from 19 to 26 (MEAN: 22.5, STD: 1.80) participate in the experiments. All participants are right-handed students with normal or corrected-to-normal vision and normal hearing. They are selected through the Eysenck Personality Questionnaire (EPQ) [15], a widely used questionnaire developed by Eysenck *et.al.* to assess an individual's personality traits. Eysenck initially conceptualized personality as several biologically-based independent dimensions of temperament: E (Extraversion/Introversion), N (Neuroticism/Stability), P (Psychoticism/Socialisation), and L (Lie/Social Desirability). Previous research has demonstrated that individuals with extroverted characteristics are prone to perform better in perceiving emotions during experiments compared to those without such characteristics [62], and people with high extraversion possess more empathy ability [40, 46]. Therefore, we rank the volunteers according to E values and selected those with high E values to participate in the experiments. This approach is adopted to ensure that participants possess the desired characteristics for accurate emotion recognition.

3.3 Protocol

In order to ensure data quality, the experiments are conducted in a controlled laboratory environment to minimize noise and other environmental disturbances. Also, the experiments are scheduled during the morning or early afternoon to avoid any confounding effects of fatigue. EEG and eye movement signals are simultaneously collected by the 62-channel electrode cap with the international 10-20 system and Tobii Pro Fusion eye tracker, respectively. EEG signals are acquired using ESI NeuroScan System at a sampling rate of 1000 Hz while eye movement signals are sampled at 250 Hz.

All subjects undergo four experimental sessions as Figure 1 shows. There are twenty trials in each session, each trial consists of two parts, where the first part is watching videos and the second part is self-assessment where subjects score their emotional intensity level from 1 to 10 points. For each session, there are only

five out of seven emotions to be elicited, which reduces the impact of subjects switching into too many emotional states. Prior to and following the presentation of each video clip, there is a 3-second countdown to alert participants to the imminent start or end of the video. The sequence of the video clips is carefully arranged to avoid sudden shifts in emotional valence, as human emotions tend to transform gradually. Eighty video clips of four sessions in total are divided into four folds, each fold contains five clips from each session and all emotional videos are equal in number. At the end of each session, participants are instructed to review all twenty video clips, recalling the emotional responses they experienced during the session and assigning continuous labels to the session's entirety via a mouse wheel. The continuous labels range between 0% and 100%, where larger values correspond to stronger elicited emotions. This study was approved by the Scientific & Technical Ethics Committee of Shanghai Jiao Tong University. All subjects are informed of the experimental process before the first session and signed up for an informed consent.

4 METHOD

4.1 Multimodal Adaptive Emotion Transformer

The overall architecture of the Multimodal Adaptive Emotion Transformer is illustrated in Figure 2. The training procedure has two steps. It is first trained using both EEG and eye movement features to endow it with the ability to process multimodal inputs. Afterwards, the backbone of MAET is frozen and we introduce the emotional prompt tuning to only tune the emotional prompts and the classifier of a single modality. Once MAET is trained, it can take either EEG or eye movements or both EEG and eye movements as input. Given an input feature $x \in \mathbb{R}^d$, where d is the dimension of the feature, x is first passed to the multi-view embedding module to map the single feature to multiple tokens from different views. Then, it is fed into the adaptive Transformer and the mixture Transformer, and finally predicts emotions by the classifiers.

4.1.1 Multi-view Embedding Module. The multi-view embedding module takes the input feature and transforms it into multiple embeddings, which aims to encourage the model to concentrate on different views of the feature. The input feature x is first transformed to v embeddings by v parallel linear layers:

$$e_i = \text{Linear}_i(x),\ i = 1, .., v \tag{1}$$

where $e_i \in \mathbb{R}^{d_e}$ and d_e is the dimension of embeddings. Another linear layer followed by an activation function is used to gate the embeddings with useful information for emotion recognition

$$\hat{e} = \sigma(\text{Linear}(x)), \tag{2}$$

where $\hat{e} \in \mathbb{R}^{d_e}$ and σ is the sigmoid function constraining the output value between 0 and 1. e_i and $\hat{e}$ are multiplied element-wise and then stacked over v embeddings, resulting in $E = (E_1, ..., E_v) \in \mathbb{R}^{v \times d_e}$. The final output can be calculated as

$$E = \text{BN}(\text{stack}(\hat{e} \odot e_i)),\ i = 1, ..., v \tag{3}$$

where $\odot$ represents Hadamard product and BN denotes batch normalization. By this means, an input feature x is converted into a sequence of tokens from different views which can be further processed by subsequent Transformer layers.

It is worthwhile to mention that the multi-view embedding module is optional for EEG because EEG features are naturally a sequence formed by multiple channels or multiple frequency bands, which can be applied directly by the multi-head self-attention. Whereas, we still adopt this module for EEG in this paper since we observe a performance boost with it incorporated.

4.1.2 Adaptive Transformer and Mixture Transformer. The adaptive Transformer and mixture Transformer are flexible components that are inspired by the mixture-of-experts Transformer [4]. These two modules are capable to cover arbitrary scenarios, such as inputting EEG only, inputting eye movements only, and inputting both EEG and eye movements, owing to the flexibility of the multi-head self-attention.

Before passing to the adaptive Transformer, the embeddings E are first prepended by a learnable class token $E_{cls} \in \mathbb{R}^{d^e}$, the function of which is to aggregate information from the whole sequence and used for emotion classification later. To incorporate the positional and modal information, learnable positional embedding $E_{pos} \in \mathbb{R}^{(v+1) \times d_e}$ and modality embedding $E_{mod} \in \mathbb{R}^{d_e}$ are added to the input embeddings, which can be formulated as

$$\tilde{E} = (E_{cls}, E_1, ..., E_v) + E_{pos} + E_{mod}, \tag{4}$$

where $\tilde{E} \in \mathbb{R}^{(v+1) \times d_e}$.

The core component in the adaptive Transformer and mixture Transformer is the same, i.e., multi-head self-attention (MHSA) [54]. The embedding $\tilde{E}$ is transformed to queries Q_i, keys K_i, and values V_i by three linear layers. The self-attention can be calculated as

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}(\frac{Q_i K_i^T}{\sqrt{d_e}})V_i. \tag{5}$$

We employ h heads for self-attention and each head can be denoted by $H_i = \text{Attention}(Q_i, K_i, V_i)$. The output of multi-head attention is $\text{Concat}(H_1, H_2, ..., H_h)W$, where W is the weight.

The adaptive Transformer introduces two modality experts to substitute the standard feed-forward network (FFN), i.e., EEG-FFN and EYE-FFN, and it adaptively selects an expert to capture the modality-specific information according to the input modality. For example, if the input is EEG-only (eye movements-only), we employ the expert of EEG-FFN (EYE-FFN) to encode the features. If the input contains multiple modalities, the EEG expert and the eye movement expert are used to process the respective modality features parallelly. The mixture Transformer just follows the vanilla Transformer, of which the Mix-FFN is expected to capture more modality interaction. We stack L_a adaptive Transformer blocks and L_m mixture Transformer blocks.

4.1.3 Fusion and Classifiers. Let $H_{cls} \in \mathbb{R}^{d_e}$ denote the class token of the mixture Transformer output. We introduce an attention-based fusion to adaptively fuse the features from multiple modalities. We first calculate the attention weights μ^{eeg} and μ^{eye} by

$$\mu^{eeg}, \mu^{eye} = \text{softmax}(\langle H_{cls}^{eeg}, W^A \rangle, \langle H_{cls}^{eye}, W^A \rangle), \tag{6}$$

where $W^A \in \mathbb{R}^{d_e}$ and $\langle , \rangle$ means dot product. Thus, the fused features are extracted by

$$H = \mu^{eeg} H_{cls}^{eeg} + \mu^{eye} H_{cls}^{eye}. \tag{7}$$

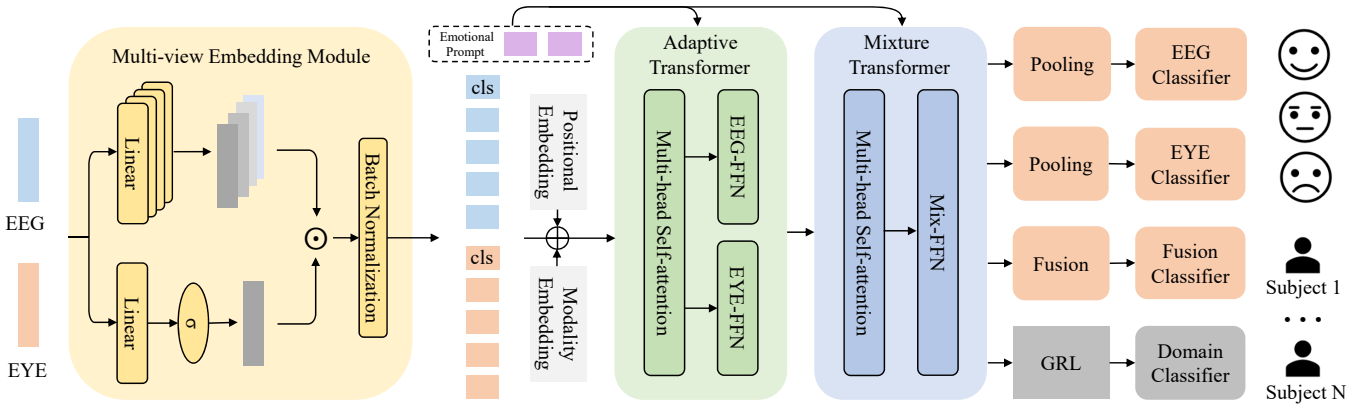


Figure 2: The architecture of MAET. MAET is a general and flexible framework for EEG and eye movements, composed of a multi-view embedding module, an adaptive Transformer block, a mixture Transformer block, and several classifiers.

Finally, a classifier that consists of a linear layer is applied to the fused features to obtain the final prediction y. The whole procedure can be formulated as

$$y^m = C_f(\mathcal{F}(x^{eeg}, x^{eye})), \tag{8}$$

where $\mathcal{F}$ represents the feature extractor of MAET, i.e., the components excluding classifiers, and C_f denotes the fusion classifier. The objective function is the cross-entropy loss:

$$\mathcal{L}_m = -\sum_{i=1}^{N} \hat{y}_i \log y_i^m, \tag{9}$$

where $\hat{y}$ is the ground truth label.

4.1.4 Emotional Prompt Tuning. We introduce emotional prompt tuning (EPT), which is inspired by the advent of prompt tuning [24, 32], to tune the model that has been trained on multimodal inputs to adapt a single modality. The idea is quite straightforward. We prepend a small set of learnable embeddings $P_i \in \mathbb{R}^{p \times d_e}$ to the feature embeddings in each Transformer layer, which are referred to as emotional prompts. The emotional prompt tuning can be formulated as

$$[\tilde{E}_{i+1}, \_] = TL_i([\tilde{E}_i, P_i]), \tag{10}$$

where TL_i denote i-th Transformer layer and $\tilde{E}_i$ denotes the feature embeddings of i-th layer. $\tilde{E}_{i+1}$ is the output as well as the input of $i+1$-th Transformer layer. After all the adaptive and mixture Transformer layers, we adopt the mean pooling over all the EEG or eye movement embeddings, followed by the classifier C_{eeg} for EEG or C_{eye} for eye movements. In this stage, we only tune the emotional prompts together with the classifier and keep the entire backbone trained on multimodal signals frozen. Thus, the ability to cope with multimodal inputs is preserved, while it learns to predict emotions using a single modality.

4.1.5 Domain Adversarial Training for Domain Generalization. EEG signals vary considerably across different subjects, which leads to the degraded generalizability of deep learning models and make cross-subject emotion recognition challenging. In order to reduce the negative impact of individual discrepancy, we exploit the adversarial domain generalization method to make the model more robust [17]. The core idea is to encourage the model to learn domain-invariant representations. Assume that for an input feature x, its corresponding domain label is d from K domains. We devise a domain classifier C_d which consists of two linear layers and the Gaussian error linear units (GELU) [20] function in between them. The domain classifier is trained jointly with other components in MAET to discriminate which domain the input belongs to. However, over-confident domain classifiers and domain label noise can lead to instability in domain adversarial training. To overcome this challenge, we adopt the environment label smoothing (ELS) [59] which encourages the domain classifier to output soft probability. For an domain label $d \in [0, 1]^K$, we transform it to $\hat{d}$ as follows

$$\hat{d}(i) = \begin{cases} \gamma, & \text{for } d(i) = 1; \\ \frac{1-\gamma}{K-1}, & \text{otherwise,} \end{cases} \tag{11}$$

where i is from 1 to K and $\sum_{i=1}^{K} \hat{d}(i) = 1$. γ is the tradeoff that controls the algorithm convergence and adversarial divergence minimization. We follow the annealing strategy [59] that gradually decreases γ during training as $\gamma = 1 - \frac{K-1}{K}\frac{t}{T}$, where t is the current training step and T is the total steps. Therefore, the loss of the domain classifier is

$$\mathcal{L}_d = -\sum_{i=1}^{N} \hat{d}_i \log C_d(\mathcal{F}(x_i)). \tag{12}$$

In order to confuse the domain classifier so that the feature extractor can learn domain-invariant representations, we introduce a gradient reverse layer (GRL) [17] which can be ignored during forward propagation and reverses the gradient passed backward from C_d to $\mathcal{F}$. Consequently, the total loss for EEG-based cross-subject emotion recognition is

$$\mathcal{L} = \mathcal{L}_{eeg} - \lambda \mathcal{L}_d, \tag{13}$$

where $\mathcal{L}_{eeg}$ is the cross-entropy loss for the EEG classifier and λ is a scaling factor that gradually changes from 0 to 1. It is suggested that $\lambda = \frac{2}{1+e^{-10t/T}} - 1$ and this strategy makes the domain classifier insensitive to noise at the early stages of the training procedure.

Table 1: Performance (accuracies and F1 scores, %) of different methods using unimodality.

Method	EEG				eye movements			
	ACC	STD	F1	STD	ACC	STD	F1	STD
KNN [9]	36.43	**05.38**	34.08	**05.79**	36.01	**06.30**	34.74	**06.33**
HCNN [31]	52.42	06.47	49.02	06.80	-	-	-	-
RGNN [63]	48.50	06.83	45.32	07.20	-	-	-	-
Transformer [54]	56.04	07.82	53.35	08.30	-	-	-	-
GCNCA[26]	58.04	07.78	**55.48**	08.30	-	-	-	-
MAET (w/o EPT)	57.84	08.80	54.94	09.41	49.99	07.17	46.72	07.94
MAET	**58.11**	08.78	54.98	09.45	**50.31**	07.14	**47.10**	07.84

4.2 Feature Extraction

4.2.1 EEG Features. Contaminated by environmental and physiological artifacts, the raw EEG signals collected during experiments contain non-negligible noise which hinders the precise analysis of brain activity. To mitigate the impact of noise, we first visually inspect the EEG signals and interpolate any bad channels using the MNE-Python toolbox [18]. We then apply a bandpass filter with cutoff frequencies of 0.1 Hz and 70 Hz to remove low-frequency noise and power-line interference. Additionally, a notch filter with a cutoff frequency of 50 Hz is applied. To reduce computational complexity, we downsample the raw EEG signals from the original sampling rate of 1000 Hz to 200 Hz.

For EEG features, differential entropy (DE) has been proven to be the most effective feature for emotion recognition, as it has a balanced ability to discriminate EEG patterns between low- and high-frequency energy [11]. We use a 256-point Short-Time Fourier Transform (STFT) with a non-overlapped Hanning window of 4 seconds to calculate the frequency domain features. The DE features are extracted in five frequency bands (delta: 1-4 Hz, theta: 4-8 Hz, alpha: 8-14 Hz, beta: 14-31 Hz, gamma: 31-49 Hz), which are defined as

$$h(X) = -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx \quad (14)$$

$$= \frac{1}{2}\log(2\pi e\sigma^2), \quad (15)$$

where the random variable X obeys Gaussian distribution $N(\mu, \sigma)$. DE is equivalent to the logarithmic energy spectrum for a fixed-length sequence in a specific band. For 62-channel EEG signals, a sample of the DE feature in the 5 frequency bands is 310-dimension.

Based on the assumption that the emotional state is defined in a continuous space and that emotional states change gradually over time, we exploit the linear dynamic system (LDS) to filter out components that are not associated with emotional states [48].

4.2.2 Eye Movement Features. For eye movement signals, various parameters can be captured by the eye tracker, such as pupil diameters, fixation details, saccade details, gaze point details, etc. Among them, pupil diameters have been demonstrated to play a critical role in emotion recognition [6]. Nonetheless, pupil diameters are highly sensitive to environmental luminance. We first employ linear interpolation to replace the missing pupil diameter samples due to eye blinking. Based on the observation that the responses of subjects to the same video in the controlled lighting environment have similar patterns, principal component analysis (PCA) is used to eliminate the effect of luminance on pupil diameters [51]. The original data are subtracted by the light reflex which is estimated by the first principal component of the observation matrix containing pupil diameter data of the same video clip from all subjects. After that, the residual part contains the emotion-associated-only pupil response. The DE features are then computed for the left and right pupil diameters using STFT in the four frequency bands with a non-overlapped Hanning window of 4 seconds. In addition to the DE features, the mean and the standard deviation of pupil diameters are also calculated. Except for pupil diameter, 21 other features are also extracted [61]. Consequently, the total number of features obtained from eye movement signals is 33.

Table 2: The accuracies and F1 scores (%) of different method using multimodality.

Method	ACC	STD	F1	STD
KNN [9]	40.44	**06.30**	37.94	**06.54**
BDAE [34]	61.55	08.74	59.11	08.87
ETF [56]	65.30	08.55	63.13	08.88
VigilanceNet [8]	62.93	07.12	60.46	07.81
MAET	**71.28**	07.74	**69.16**	08.35

5 EXPERIMENTAL RESULTS

5.1 Implementation Details

For the hyperparameters of MAET, the number of adaptive Transformer blocks L_a and mixture Transformer blocks L_m is set to 2 and 1, respectively. We empirically set the number of views $v = 5$ in the multi-view embedding module. The number of heads h is 4 in MHSA. The embedding dimension d_e is tuned from $\{32, 64\}$. The batch size is 64 in subject-dependent experiments and 256 in cross-subject experiments. We use AdamW [35] as the optimizer with the learning rate tuned from $\{0.00003, 0.0001, 0.0003\}$. Moreover, we tune the weight decay from $\{0.0001, 0.01, 0.1\}$. The prompt length p is 1 or 2. Note that the domain adversarial training is only employed in the cross-subject conditions. The emotional prompt tuning is only employed in Section 5.2. Otherwise, MAET is directly trained using solely EEG features with the cross-entropy loss.

5.2 Unimodal and Multimodal Emotion Recognition

Aiming to evaluate the efficacy of EEG and eye movements in identifying emotions, we construct a subject-dependent model for each subject. Specifically, we merge the data from all four sessions of one subject and then partition the data into four folds for carrying out a four-fold cross-validation. Notably, the input EEG and eye movement features are transformed by z-score normalization.

5.2.1 Classification Performance of EEG. The classification performance of six baseline classifiers and our newly developed MAET are compared systematically, K nearest neighbor (KNN) [9] (K is set to 1), hierarchical convolutional neural network (HCNN) [31], regularized graph neural network (RGNN) [63], Transformer [54], and graph convolutional network with channel attention (GCNCA) [26]. Note that the EmotionDL proposed in RGNN is not implemented in this paper. All methods are implemented strictly under the same conditions and are compared with each other fairly. Table 1 shows the average accuracies and weighted F1 scores for each method.

For classifying the seven emotions, MAET attains the most accurate discriminating capacity. Specifically, the highest prediction accuracy of 58.11% is acquired by the MAET machine while utilizing the total band, highlighting the effectiveness of MAET. Figure 3 (a) depicts the confusion matrix of MAET using EEG signals exclusively. It can be observed that the surprise and fear emotions are distinguished by MAET with higher accuracy than other emotions. In addition, the happy emotion is prone to be misclassified as surprise while neutral is more likely to be confused with sadness. Besides, compared to other emotions, the sad and angry emotions display a greater likelihood of being misclassified as each other in the classification based on EEG signals, which indicates the similarity between the neural patterns of sad and angry emotions.

5.2.2 Classification Performance of Eye Movements. For eye movements, we compare our proposed MAET with KNN since other baseline methods are unable to handle eye movement features. The results are shown in Table 1. It is conspicuous that MAET obtains the highest prediction accuracy of 50.31% and the highest F1 score of 47.10%. Figure 3 (b) presents the confusion matrix of MAET using eye movement signals alone. It is evident from the table that eye movement signals exhibit remarkable performance in distinguishing neutral and fear emotions. In spite of this, isolate eye movement signals have relatively poor performance in classifying happy and disgust emotions, whose accuracies are lower than 40%. Observed from Figure 3 (a) and (b), EEG signals attain better results in discriminating happiness, surprise, disgust, and anger while eye movement signals acquire more accurate results in classifying neutral, sadness, and fear. It is worth noting that there are some similar eye movement patterns between happy and angry emotions for the reason that 27.53% happy emotions have been recognized as anger.

In addition, the second line from the bottom in Table 1 presents the results of only tuning the classifiers. The decreased performance highlights the effectiveness of emotional prompt tuning.

5.2.3 Classification Performance of Multimodal signals. Table 2 displays the results of different models using both EEG and eye movements. A systematic comparison is conducted between KNN, Bimodal Deep AutoEncoder (BDAE) [34], Emotion Transformer Fusion (ETF) [56], VigilanceNet [8], and MAET. For KNN, the EEG features and eye movement features are directly concatenated into 343-dimensional features. MAET outperforms the other methods with the best accuracy of 71.28% and F1 score of 69.16%, which illustrates its effectiveness. The confusion matrix of MAET using multimodal signals is shown in Figure 3 (c). It could be viewed that MAET acquires outstanding accuracy on the task of classifying surprise, neutral, fear, and anger emotions. Among all emotions, the accuracy of distinguishing fear emotion is the highest at 82.41%. It is evident that most emotions can be classified more accurately while using multimodality than using EEG or eye movement signals individually. These results demonstrate that multimodality can significantly improve classification performance, which indicates the complementary properties between EEG and eye movements.

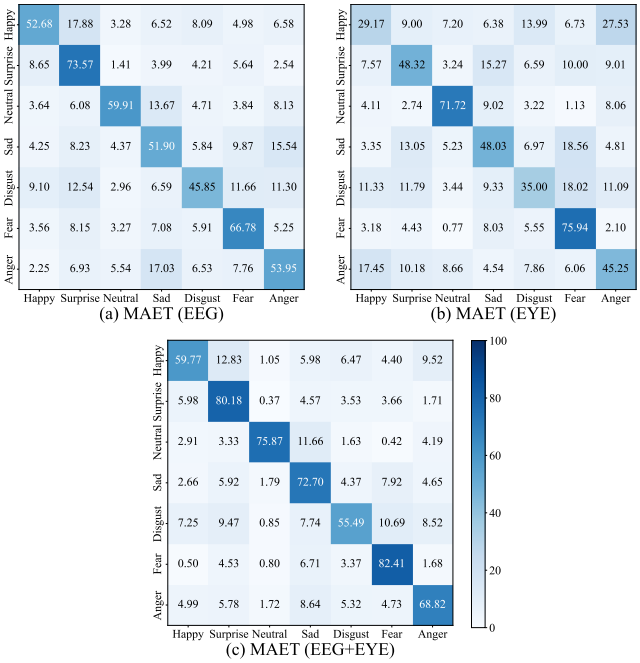


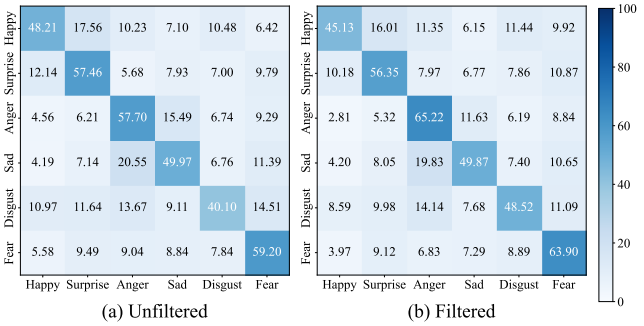
Figure 3: Confusion matrices of MAET using EEG and eye movements or both. The horizontal axis represents the predicted labels and the vertical axis represents the true labels.

Table 3: The performance (%) of different methods of cross-subject experiments.

Method	ACC	STD	F1	STD
KNN [9]	20.85	04.56	20.23	04.49
HCNN [31]	39.88	04.94	38.18	05.06
RGNN [63]	37.49	05.44	34.52	04.83
Transformer [54]	40.36	05.22	37.76	05.54
GCNCA [26]	38.68	**03.94**	37.25	**03.65**
MAET (w/o AT)	40.69	05.50	38.47	06.09
MAET	**40.90**	05.52	**38.85**	06.07

Table 4: The accuracies and F1 scores (%) of different methods with/without the filtered high-induced data.

Method	Unfiltered		Filtered	
	ACC	F1	ACC	F1
KNN [9]	31.39	29.92	33.98	32.86
HCNN [31]	46.45	44.29	50.24	45.78
RGNN [63]	43.61	42.51	50.70	49.55
Transformer [54]	49.71	48.71	56.84	56.19
GCNCA [26]	52.74	52.21	58.15	**58.16**
MAET	**52.86**	**52.26**	**58.24**	58.08



(a) Unfiltered (b) Filtered

Figure 4: Confusion matrices of MAET with/without the filtered high-induced data using continuous labels.

5.3 Cross-subject Emotion Recognition

One of the essential questions is whether it is reliable and robust in recognizing emotions of a new subject, whose physiological signals have never been recorded and fed into the classifiers. The strategy we adopt for measuring the cross-subject performance is leave-one-subject-out (LOSO) cross-validation. The results are depicted in Table 3. Due to the variability between distinct subjects, the performance of all methods is singularly poor compared with that in subject-dependent conditions and the performance degradation is nearly 20%. MAET acquires the highest accuracy of 40.90% and F1 score of 38.85%, demonstrating the robustness of MAET. It is worth noting that MAET without adversarial training (AT) attains the second-highest accuracy of 40.69%, which implies that adversarial training is helpful for cross-subject situations to some extent.

5.4 Analysis of Continuous Labels

The intensity score associated with an emotion is a crucial indicator of the elicitation level and the quality of the collected physiological data. Several previous studies [49] have studied emotions by continuous labels which can measure affective arousal sensitively. Due to the ambiguous definition of intensity score for neutrality, we exclude neutral emotion from our analysis in this experiment.

To further investigate the relevance between classification performance and the intensity level, we conduct a comparison of the classification performance of each method under unfiltered and filtered situations, and low-induced data are filtered under the filtered situation. The criteria for judging whether the EEG signals are high-induced is that the score of the corresponding video clip is larger than a threshold of 50%. For the purpose of discriminating the quality between high-induced and low-induced data, the smoothing algorithm LDS introduced in Section 4.2.1 is not employed in this experiment. The results are displayed in Table 4. From Table 4, we can see that the classification accuracy and F1 score increase considerably by just using the filtered data for all methods. The highest accuracy is achieved by MAET of 58.24%. The increasing rate of accuracy in the filtered case is over 5% and the greatest increment is obtained by Transformer of 7.13%.

Figure 4 depicts the confusion matrix of MAET under unfiltered and filtered situations. It is observed that by filtering high-induced data, the accuracy of the anger, disgust, and fear emotion rises 7.52%, 8.42%, and 4.7%, which underscores the importance of filtering for discriminating these three emotions. However, there is a slight cutback in the classification accuracy of the happy, surprise, and sad emotions. The reason might be that for the happy, surprise, and sad emotions, it takes a longer time for subjects to be evoked by stimuli materials, which results in the lack of physiological data after filtering. As deep models require large amounts of data, inadequate data after filtering may account for the decrease in the accuracies of these three emotions. This observation further demonstrates the effectiveness of filtering for happy, surprise, and sad emotions. The finding suggests that filtered high-induced data is significant for classifying easy-evoked emotions.

6 CONCLUSION

In this study, we have developed a novel multimodal emotion dataset for the seven basic emotions (happy, sad, fear, disgust, surprise, anger, and neutral) with EEG and eye movement signals. Besides, our dataset possesses continuous labels which indicate the affective intensity level subjects experienced during watching videos. 80 video clips are chosen as our stimuli materials and 20 subjects are recruited by the EPQ questionnaire in our experiments. Besides, We have proposed a novel method MAET which is capable to deal with unimodal and multimodal inputs flexibly. The performance of different methods has been evaluated systematically in unimodal and multimodal cases. Furthermore, we have conducted the cross-subject experiment, using the LOSO cross-validation to examine the performance of each method. On the other hand, MAET provides a general baseline for future research. Moreover, a comparison between the unfiltered and filtered situations has been carried out to explore the effect of filtering data according to continuous labels. The experimental results show a considerable accuracy increment with the filtered data.

ACKNOWLEDGMENTS

This work was supported in part by grants from National Natural Science Foundation of China (Grant No. 61976135), STI 2030-Major Projects+2022ZD0208500, Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

REFERENCES

[1] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56–76.

[2] Soraia M. Alarcão and Manuel J. Fonseca. 2019. Emotions Recognition Using EEG Signals: A Survey. *IEEE Transactions on Affective Computing* 10, 3 (2019), 374–393. https://doi.org/10.1109/TAFFC.2017.2714671

[3] Arjun Arjun, Aniket Singh Rajpoot, and Mahesh Raveendranatha Panicker. 2021. Introducing attention mechanism for EEG signals: Emotion recognition with vision transformers. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 5723–5726.

[4] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. 2021. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358* (2021).

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

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

[7] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences* 582 (2022), 593–617. https://doi.org/10.1016/j.ins.2021.10.005

[8] Xinyu Cheng, Wei Wei, Changde Du, Shuang Qiu, Sanli Tian, Xiaojun Ma, and Huiguang He. 2022. VigilanceNet: Decouple Intra-and Inter-Modality Learning for Multimodal Vigilance Estimation in RSVP-Based BCI. In *Proceedings of the 30th ACM International Conference on Multimedia*. 209–217.

[9] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27.

[10] Fernando Lopes da Silva. 2013. EEG and MEG: relevance to neuroscience. *Neuron* 80, 5 (2013), 1112–1128.

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

[12] Monika Dubey and Lokesh Singh. 2016. Automatic emotion recognition using facial expression: a review. *International Research Journal of Engineering and Technology (IRJET)* 3, 2 (2016), 488–492.

[13] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124.

[14] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

[15] Sybil BG Eysenck, Hans J Eysenck, and Paul Barrett. 1985. A revised version of the psychoticism scale. *Personality and Individual Differences* 6, 1 (1985), 21–29.

[16] N. Fragopanagos and J.G. Taylor. 2005. Emotion recognition in human–computer interaction. *Neural Networks* 18, 4 (2005), 389–405. https://doi.org/10.1016/j.neunet.2005.03.006

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[18] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. 2013. MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience* 7, 267 (2013), 1–13. https://doi.org/10.3389/fnins.2013.00267

[19] Kairui Guo, Rifai Chai, Henry Candra, Ying Guo, Rong Song, Hung Nguyen, and Steven Su. 2019. A hybrid fuzzy cognitive map/support vector machine approach for EEG-based emotion classification using compressed sensing. *International Journal of Fuzzy Systems* 21 (2019), 263–273.

[20] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[21] Yu-Liang Hsu, Jeen-Shing Wang, Wei-Chun Chiang, and Chien-Han Hung. 2017. Automatic ECG-based emotion recognition in music listening. *IEEE Transactions on Affective Computing* 11, 1 (2017), 85–99.

[22] Xin Hu, Fei Wang, and Dan Zhang. 2022. Similar brains blend emotion in similar ways: Neural representations of individual difference in emotion profiles. *Neuroimage* 247 (2022), 118819.

[23] S Jerritta, M Murugappan, Khairunizam Wan, and Sazali Yaacob. 2014. Emotion recognition from facial EMG signals using higher order statistics and principal component analysis. *Journal of the Chinese Institute of Engineers* 37, 3 (2014), 385–394.

[24] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 709–727.

[25] Ziyu Jia, Youfang Lin, Jing Wang, Zhiyang Feng, Xiangheng Xie, and Caijie Chen. 2021. HetEmotionNet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1047–1056.

[26] Wei-Bang Jiang, Li-Ming Zhao, Ping Guo, and Bao-Liang Lu. 2021. Discriminating surprise and anger from EEG and eye movements with a graph network. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1353–1357.

[27] Stamos Katsigiannis and Naeem Ramzan. 2017. 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* 22, 1 (2017), 98–107.

[28] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* 7 (2019), 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

[29] Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 12 (2008), 2067–2083.

[30] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2011), 18–31.

[31] Jinpeng Li, Zhaoxiang Zhang, and Huiguang He. 2018. Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cognitive Computation* 10 (2018), 368–380.

[32] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[33] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. 2021. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems* 14, 2 (2021), 715–729.

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

[35] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[36] Iris B. Mauss and Michael D. Robinson. 2009. Measures of emotion: A review. *Cognition and Emotion* 23, 2 (2009), 209–237. https://doi.org/10.1080/02699930802204677

[37] RAM KUMAR Mdupu, CHIRANJEEVI Kothapalli, VASANTHI Yarra, SONTI Harika, and Cmak ZEELAN Basha. 2020. Automatic Human Emotion Recognition System using Facial Expressions with Convolution Neural Network. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 1179–1183. https://doi.org/10.1109/ICECA49313.2020.9297483

[38] Wafa Mellouk and Wahida Handouzi. 2020. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science* 175 (2020), 689–694.

[39] Shraddha A. Mithbavkar and Milind S. Shah. 2021. Analysis of EMG Based Emotion Recognition for Multiple People and Emotions. In *2021 IEEE 3rd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*. 1–4. https://doi.org/10.1109/ECBIOS51820.2021.9510858

[40] Hye Jeong Park and Jae Hwa Lee. 2020. Looking into the personality traits to enhance empathy ability: A review of literature. In *HCI International 2020-Posters: 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*. Springer, 173–180.

[41] R.W. Picard, E. Vyzas, and J. Healey. 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10 (2001), 1175–1191. https://doi.org/10.1109/34.954607

[42] Aniket Singh Rajpoot, Mahesh Raveendranatha Panicker, et al. 2022. Subject independent emotion recognition using EEG signals employing attention driven neural networks. *Biomedical Signal Processing and Control* 75 (2022), 103547.

[43] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[44] Heini Saarimäki, Athanasios Gotsopoulos, Iiro P. Jääskeläinen, Jouko Lampinen, Patrik Vuilleumier, Riitta Hari, Mikko Sams, and Lauri Nummenmaa. 2015. Discrete Neural Signatures of Basic Emotions. *Cerebral Cortex* 26, 6 (04 2015), 2563–2573. https://doi.org/10.1093/cercor/bhv086 arXiv:https://academic.oup.com/cercor/article-pdf/26/6/2563/17309892/bhv086.pdf

[45] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, and Gholamreza Anbarjafari. 2019. Emotion recognition from skeletal movements. *Entropy* 21, 7 (2019), 646.

[46] Teresa Schreckenbach, Falk Ochsendorf, Jasmina Sterz, Miriam Rüsseler, Wolf Otto Bechstein, Bernd Bender, and Myriam N Bechtoldt. 2018. Emotion recognition and extraversion of medical students interact to predict their empathic communication perceived by simulated patients. *BMC medical education*

18, 1 (2018), 1–10.

[47] Zhijuan Shen, Jun Cheng, Xiping Hu, and Qian Dong. 2019. Emotion Recognition Based on Multi-View Body Gestures. In *2019 IEEE International Conference on Image Processing (ICIP)*. 3317–3321. https://doi.org/10.1109/ICIP.2019.8803460

[48] Li-Chen Shi and Bao-Liang Lu. 2010. 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*. 6587–6590. https://doi.org/10.1109/IEMBS.2010.5627125

[49] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 2016. Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions on Affective Computing* 7, 1 (2016), 17–28. https://doi.org/10.1109/TAFFC.2015.2436926

[50] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2011), 42–55.

[51] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing* 3, 2 (2012), 211–223. https://doi.org/10.1109/T-AFFC.2011.37

[52] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 2020. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing* 11, 3 (2020), 532–541. https://doi.org/10.1109/TAFFC.2018.2817622

[53] Bo Sun, Liandong Li, Xuewen Wu, Tian Zuo, Ying Chen, Guoyan Zhou, Jun He, and Xiaoming Zhu. 2016. Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild. *Journal on Multimodal User Interfaces* 10 (2016), 125–137.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[55] Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L Crowley. 2022. Emotion Recognition with Pre-Trained Transformers Using Multimodal Signals. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[56] Yiting Wang, Wei-Bang Jiang, Rui Li, and Bao-Liang Lu. 2021. Emotion transformer fusion: Complementary representation properties of EEG and eye movements on recognizing anger and surprise. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1575–1578.

[57] Zhe Wang, Yongxiong Wang, Chuanfei Hu, Zhong Yin, and Yu Song. 2022. Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model. *IEEE Sensors Journal* 22, 5 (2022), 4359–4368.

[58] D. Watts, R. F. Pulice, J. Reilly, A. R. Brunoni, F. Kapczinski, and I. C. Passos. 2022. Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Transl Psychiatry* 12, 1 (2022), 332. https://doi.org/10.1038/s41398-022-02064-z

[59] YiFan Zhang, Xue Wang, Jian Liang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2023. Free Lunch for Domain Adversarial Training: Environment Label Smoothing. In *International Conference on Learning Representations*.

[60] Li-Ming Zhao, Rui Li, Wei-Long Zheng, and Bao-Liang Lu. 2019. Classification of five emotions from EEG and eye movement signals: complementary representation properties. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 611–614.

[61] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics* 49, 3 (2018), 1110–1122.

[62] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development* 7, 3 (2015), 162–175.

[63] Peixiang Zhong, Di Wang, and Chunyan Miao. 2022. EEG-Based Emotion Recognition Using Regularized Graph Neural Networks. *IEEE Transactions on Affective Computing* 13, 3 (2022), 1290–1301. https://doi.org/10.1109/TAFFC.2020.2994159