SEED-VII: A Multimodal Dataset of Six Basic Emotions With Continuous Labels for Emotion Recognition

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, *Member, IEEE*, and Bao-Liang Lu, *Fellow, IEEE*

***Abstract*—Recognizing emotions from physiological signals is a topic that has garnered widespread interest, and research continues to develop novel techniques for perceiving emotions. However, the emergence of deep learning has highlighted the need for comprehensive and high-quality emotional datasets that enable the accurate decoding of human emotions. To systematically explore human emotions, we develop a multimodal dataset consisting of six basic (happiness, sadness, fear, disgust, surprise, and anger) emotions and the neutral emotion, named SEED-VII. This multimodal dataset includes electroencephalography (EEG) and eye movement signals. The seven emotions in SEED-VII are elicited by 80 different videos and fully investigated with continuous labels that indicate the intensity levels 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 the MAET to mitigate subject discrepancies, which enhances domain generalization. Our extensive experiments, encompassing both subject-dependent and cross-subject conditions, demonstrate the superior performance of the MAET in terms of handling various inputs. Continuous labels are used to filter the data with high emotional intensity, and this strategy is proven to be effective for attaining improved emotion recognition performance. Furthermore, complementary properties between the EEG signals and eye movements and stable neural patterns of the seven emotions are observed.**

Received 21 December 2023; revised 14 August 2024; accepted 18 October 2024. Date of publication 23 October 2024; date of current version 27 May 2025. This work was supported in part by STI 2030-Major Projects under Grant +2022ZD0208500, in part by the National Natural Science Foundation of China under Grant 62376158, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZD ZX, in part by Shanghai Pujiang Program under Grant 22PJ1408600, in part by the Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program under Grant YG2023ZD25, Grant YG2024ZD25, and Grant YG2024QNA03, in part by Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University under Grant 21TQ1400203 and in part by GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine. Recommended for acceptance by W. Wu. *(Wei-Bang Jiang and Xuan-Hao Liu contributed equally to this research. (Corresponding author: Bao-Liang Lu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Review Committee for Human-Related Scientific and Technological Research of Shanghai Jiao Tong University under Application No. I2022283I, and performed in line with the Multimodal Emotion Recognition Experiment Based on EEG and Eye Movements.

The authors are with the Center for Brain-Like Computing and Machine Intelligence, Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, and Brain Science and Technology Research Center, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: 935963004@sjtu.edu.cn; haogram_sjtu@sjtu.edu.cn; weilong@sjtu.edu.cn; bllu@sjtu.edu.cn).

Digital Object Identifier 10.1109/TAFFC.2024.3485057

***Index Terms*—Basic emotions, continuous label, EEG, emotion recognition, eye movements, multimodal dataset.**

I. Introduction

EMOTION recognition plays a crucial role in developing emotional artificial intelligence systems [1], [2] and affective brain-computer interfaces [3], [4], which enable machines to attain emotional intelligence [5], and allow computers to identify, understand, and respond to the emotions of human beings. Moreover, existing studies have revealed the strong relationship between emotions and mental illnesses. Mood states such as depression, attention deficit hyperactivity disorder, anxiety disorder, and internet addiction can be identified in patients through their emotional states [6], [7], [8], [9]. Given the complexity and importance of emotion, a psychophysiological process triggered by various factors [5], researchers in the fields of psychology, neuroscience, and computer science have been exploring emotion recognition for years [10], [11]. However, the challenges of detecting and analyzing human emotions remain largely unexplored.

In recent years, a variety of physiological and nonphysiological signals have been employed for emotion recognition [11]. Nonphysiological signals, such as speech [12], [13], [14], facial expressions [15], [16], [17], and body movements [18], [19] have been utilized by researchers to recognize human emotions. However, nonphysiological signals can be easily falsified and are thus untrustworthy, as individuals may conceal their true emotions. In contrast, physiological signals, such as electroencephalogram (EEG) [11], [20], [21], electromyogram (EMG) [22], [23], and electrocardiogram (ECG) [21], [24], [25] signals, provide more reliable and stable options than nonphysiological signals. Specifically, intramuscular EMG, involving inserting needles into the muscle to record electrical activity, is more reliable and difficult to falsify than surface EMG due to its invasive nature and the specific muscle responses it captures.

Among all the available physiological signals, EEG signals have been shown to outperform other signals such as galvanic skin response (GSR), respiration (RSP), and ECG in emotion recognition tasks [11], [20], [26]. EEG signals are inherently correlated with brain activity and have been investigated in many fields, such as psychology and neuroscience [6], [27]. In addition, eye movement signals have been proven capable of acquiring properties that are complementary to those of EEG

signals in multimodal emotion recognition scenarios [28], [29]. Therefore, we collect EEG signals, as well as eye movement signals, to create a novel multimodal dataset named SJTU Emotion EEG Dataset VII (SEED-VII), which includes six basic emotions (happiness, sadness, fear, disgust, surprise, and anger) [30] and the neutral emotion.

There are two main types of approaches for characterizing emotions: the dimensional model and the discrete categorical model. The most well-known dimensional model is the 2D spatial Russell model proposed in 1980, where all affective concepts are located at a point with valence and arousal dimensions [31]. The valence dimension represents whether the emotions are positive or negative, while the arousal dimension 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 dataset [32] and MAHNOB-HCI dataset [28]. Unlike the dimensional approach which portrays emotions continuously, the discrete categorical model, first proposed by Ekman in the 1970s, classifies emotions into a set of discrete statuses [30]. Ekman's theory identifies 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 the SEED dataset [20] and emotion research conducted using functional Magnetic Resonance Imaging (fMRI) [33]. Regarding the insufficient and incomplete study for Ekman's basic emotions in EEG-based emotion analysis, our SEED-VII dataset is based on the discrete model and examines the EEG and eye movement signals associated with seven emotions, including the six basic emotions and the neutral state.

EEG signals are convenient and noninvasive physiological data for emotion recognition due to their harmlessness, inexpensiveness, and quick acquisition [11]. However, the existing EEG emotion datasets such as MAHNOB-HCI [28], DEAP [32], and SEED [20] have limited diversity in terms of their emotion categories and short recording durations, which restricts their potential for use in data analysis and improving the performance of emotion recognition. Furthermore, neuroscience and cognitive science studies have shown that emotions are complex and dynamic physiological processes that exhibit various intensities and states over time [10]. Therefore, recording continuous intensity labels is a practical way to study these changes.

To the best of our knowledge, there are limited public EEG emotion datasets with continuous labels for representing the intensity levels of elicited emotions [34]. Additionally, multimodal signals have been proven to be effective in emotion recognition tasks [35], highlighting the need to record other physiological and nonphysiological signals during experiments. To address these issues, we develop a novel multimodal dataset with continuous labels for emotion recognition focusing on six basic emotions and the neutral emotion. Our dataset features more than 14,000 seconds of recordings, making it longer than most of the existing EEG datasets, which typically record less than 4,000 seconds of data.

To address the challenges of emotion recognition, many methods have been explored and applied in recent years, among which deep learning techniques have been certified to outperform other approaches [11]. In recent decades, deep belief networks (DBNs) [20], convolutional neural networks (CNNs) [36], graph neural networks (GNNs) [37], graph convolutional neural networks (GCNNs) [38], [39], and Transformer [40], [41] have been employed for emotion recognition and have achieved good discrimination accuracy. Utilizing the attention mechanism, Transformer can calculate the relevance inside a sequential structure [42]. Moreover, Transformer has a strong ability to capture intermodal and intramodal interactions and is suitable for multimodal emotion recognition.

Although emotion recognition can be performed more efficiently with multiple modalities, few methods have been designed specifically for multimodal cases. In this paper, we propose a novel Multimodal Adaptive Emotion Transformer (MAET) that possesses specialized modules that enable it to flexibly operate on both unimodal and multimodal inputs. MAET is first trained with EEG and eye movement features, aiming to learn how to address multimodal inputs. We subsequently leverage emotional prompt tuning to enable the MAET to recognize emotions using a single modality while still maintaining the ability to process multimodal features. Moreover, subject discrepancies are obscured by the use of adversarial training to promote domain generalization in the MAET.

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

1) We introduce a novel multimodal emotion dataset focusing on six basic emotions (happiness, sadness, fear, disgust, surprise, and anger) and the neutral emotion, with recorded EEG and eye movement signals. Additionally, continuous labels representing the intensity levels of the corresponding seven emotions are collected. The dataset is publicly available.[1]
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 proposed MAET model alleviates subject discrepancies by adopting adversarial training to improve its domain generalization capabilities.
3) We conduct systematic experiments under various conditions, including unimodal and multimodal conditions, as well with subject-dependent and subject-independent conditions, on the SEED-VII dataset to evaluate the efficiency of our MAET model compared to that of other classifiers. Moreover, we investigate not only the neural signatures and stable patterns but also the statistics of eye movement signals.
4) We analyze the effectiveness of filtering high-induced data using continuous labels. The experimental results indicate that filtering high-induced data can significantly enhance the emotion discrimination ability of the proposed approach. Moreover, we conduct a low- and high-induced classification experiment to explore the possibility of filtering without continuous labels.

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

TABLE I
A SURVEY OF THE AVAILABLE PUBLIC EMOTION EEG DATASETS USING VIDEO STIMULI MATERIALS

Dataset	#Subject/ Session	#Videos	Total time	Physiological signals	EEG signals	Continuous labels	Emotion states
MAHNOB-HCI [28]	27/27	20	1628 s	EEG, ECG, GSR, gaze, temperature, and respiration	32/256 Hz	No	Valence and arousal
DEAP [32]	32/32	40	2400 s	EEG, EOG, EMG, BVP, GSR, temperature, and respiration	32/512 Hz	No	Valence, arousal, and liking
DREAMER [21]	23/23	18	3582 s	EEG, ECG	14/128 Hz	No	Valence, arousal, and dominance
SEED [20]	15/45	15	3394 s	EEG and eye movements	62/1000 Hz	No	Positive, neutral, and negative
FACED [43]	123/123	28	1876 s	EEG	32/250 Hz	No	Anger, disgust, fear, sadness, neutrality, joy, amusement, inspiration, and tenderness
HR-EEG4EMO [44]	27/27	13	2015 s	EEG, ECG, GSR, respiration, SpO2, and pulse rate	257/1000 Hz	No	Positive and negative
MPED [26]	23/23	28	5444 s	EEG, ECG, RSP, GSR	62/1000 Hz	No	Joy, funniness, anger, disgust, fear, sadness, and neutrality
SEED-VII	20/80	80	14097 s	EEG and eye movements	62/1000 Hz	Yes	Happiness, surprise, neutrality, disgust, fear, sadness, and anger

II. RELATED WORK

A. EEG Dataset for Emotion Recognition

Given the extensive attention that the emotion recognition task using EEG signals has received, an increasing number of methods for evaluating emotional states have been proposed [11]. Hence, comprehensive and high-quality emotional datasets are urgently needed for researchers to evaluate the performance of their methods. To date, there are several datasets for classifying emotions that include EEG recordings alone or EEG recordings along with other modalities. In this section, we review several of the existing public emotion datasets generated from video stimulus materials with EEG signals. A survey of the main public datasets used in the literature is presented in Table I. We investigate the basic information of each dataset, including the number of sessions, the number of videos, the total time of these videos, the recorded physiological signals, the number of channels and recording frequency of the raw EEG signals, the availability of continuous labels, and the number of emotion states studied.

By adopting the valence arousal model, the DEAP [32] and MAHNOB-HCI [28] datasets recorded EEG signals as well as other physiological signals, such as GSR, ECG, and EMG signals, for emotion research. The naive Bayes and SVM classifiers have been used to conduct the research on DEAP and MAHNOB-HCI, respectively. DEAP revealed that EEG signals were better at predicting arousal while peripheral physiological signals were better at predicting valence. Notably, eye gaze data were proven to be the best single modality for classifying both arousal and valence based on the MAHNOB-HCI dataset [28], highlighting the potential effectiveness of eye movement signals in emotion recognition tasks. To increase the applicability of affective computing 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 [21]. An SVM classifier with a Radial Basis Function (RBF) kernel was used to discriminate low and high valence, arousal, and dominance levels in the DREAMER dataset. The datasets above are based on the valence-arousal model so they are consequently inappropriate for conducting research on particular discrete emotion states. For example, as one of the most widely used EEG emotion datasets, DEAP uses music videos as stimuli, which makes it difficult to evoke particular emotions. Hence, research conducted based on the DEAP dataset can only roughly classify high/low valence and arousal, instead of precise emotions such as happiness or sadness.

Unlike the datasets mentioned above, the SEED [20] dataset utilized a discrete model to observe the EEG and eye movement states of particular emotions. Fifteen film clips were selected to evoke positive, neutral, and negative emotions. Based on the SEED dataset, Zheng and Lu investigated critical bands and channels for EEG emotion recognition. It was found that using the EEG signals derived from channels in the lateral temporal areas with all frequency bands yielded the best classification accuracy. To acquire high-resolution EEG (HR-EEG) signals, Becker et al. selected 13 videos that consisted of 7 positive emotions and 6 negative emotions from FilmStim to obtain HR-EEG data along with other physiological signals from 27 subjects [44]. MPED [26] includes a wide range of emotions, such as joy, funniness, anger, fear, disgust, sadness, and neutrality. Multiple physiological signals were recorded with 28 emotional videos used to elicit emotions from 23 subjects. Hu et al. constructed the THU-EP dataset [45] in 2022, collecting EEG signals from 80 subjects who responded to 28 video clips consisting of nine emotions, including four positive emotions, such as joy and amusement, four negative emotions like anger and disgust. THU-EP dataset was then developed by recruiting more subjects to 123 in total to form a bigger dataset called FACED [43]. By using the discrete model, researchers can investigate precise emotion states and combinations of emotions. However, the existing datasets cannot effectively satisfy the requirements of comprehensive and high-quality emotional data for the following reasons. 1) Limited emotion state categories have been studied; 2) inadequate videos are available for

inducing each emotion state; and 3) their videos have short total times. SEED includes happy, sad, and neutral emotions while HR-EEG4EMO contains only positive and negative emotions. The THU-EP and MPED datasets involve more than 7 emotions, but at most 4 videos were chosen for each emotion, ignoring the diversity of the emotional stimuli for evoking various affctive states. The average total time of these datasets described above is only 3334 seconds.

Besides collecting the brain signals using neural image techniques, requiring subjects to rate their own emotional intensity also assists in evaluating the quality of emotion elicitation and analyzing affective states. Koide-Majima et al. [46] recruited 166 annotators who did not participated in the main experiments to obtain the emotion ratings. However, this rating method, while ensuring relatively high objective assessment of the emotional stimuli by large amount of annotators, is unable to rating the affective states of subjects who has been recorded brain signals. In our experiment, we asked each subject to rate their emotion intensity after undergoing whole collecting process each session, which reflects the accurate and true intensity of different subjects.

B. EEG-Based Emotion Recognition

As EEG signals have been proven to be the most promising physiological signals for emotion recognition, many emotion recognition algorithms based on EEG signals have been proposed over the years [11]. Zheng and Lu employed a deep belief network to investigate the critical frequency bands and channels of EEG signals for emotion recognition [20]. By reshaping and flattening EEG signals to image-like tensors according to their spatial relationships, Li et al. used a hierarchical convolutional neural network (HCNN) to learn the spatial pattern of each emotion [36]. Alhagry et al. proposed an EEG feature extraction algorithm using long short-term memory (LSTM) and applied the obtained features for classifying low/high levels of valence and arousal [47].

To better extract topographical information from EEG signals, a regularized graph neural network (RGNN), which can capture both local and global interchannel relations, was used by Zhong et al. for emotion detection [37]. Song et al. adopted a dynamic graph convolutional neural network (DGCNN) for emotion discrimination, which can dynamically learn the intrinsic relationships between EEG channels [39]. Jiang et al. proposed a graph convolutional network with channel attention (GCNCA) to classify angry and surprised emotions [38].

Recently, Transformer has been used for emotion recognition. For example, Wang et al. proposed a Transformer-based model to hierarchically learn discriminative spatial information [40]. Utilizing an attention mechanism on raw EEG signals, Arjun et al. achieved excellent accuracy rates of 99.4% and 99.1% when classifying valence and arousal, respectively [48]. Rajpoot et al. improved upon LSTM and CNNs by using an attention mechanism for subject-independent emotion recognition, and they achieved state-of-the-art performance [49]. These excellent results demonstrate the effectiveness of attention mechanisms.

C. Multimodal Emotion Recognition

An emotion is an internal subjective experience and is always accompanied by various complex but imperceptible physiological manifestations in addition to facial expressions, such as activation in particular cerebral cortex areas [33] and pupil diameter fluctuations [50]. Hence, the application of multimodal signals can provide improved discrimination capabilities, and this approach has been widely used in emotion recognition due to the potential complementary properties of different modalities [29], [51]. However, how to effectively combine multimodal signals is still a challenging problem.

Sun et al. used a hierarchical classifier with hybrid fusion to distinguish emotions [52]. A fuzzy cognitive map and an SVM were employed by Guo et al. to form a hybrid classifier for emotion recognition [53]. A two-stream heterogeneous graph recurrent neural network was developed by Jia et al. to classify emotions. This approach can fuse spatial-spectral-temporal domain features in a unified framework [54]. Excavating and fusing information from various modalities using deep learning methods has proven to be efficient. With the invention of attention mechanisms, an increasing number of deep fusion methods have been developed based on such mechanisms. Liu et al. proposed a deep canonical correlation analysis (DCCA) approach with an attention-based fusion strategy to perform multimodal emotion recognition [35]. 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 their limited adaptability to unimodal signals. Some existing VAE-based emotion decoding methods [56] can handle both single-modal and multi-modal inputs. These models often use a shared latent space to integrate multiple modalities, extracting joint representations to improve emotion recognition accuracy. In contrast, our method mainly focuses on the fusion of different modalities and introduces modality-specific expert modules to adaptively learn inter- and intra- modality information.

III. Experimental Setup

A. Stimuli

The emotion experiments performed in this paper were designed to simultaneously record EEG and eye movement signals during the elicitation of seven emotions (happiness, sadness, disgust, fear, surprise, anger, and neutrality). The selection of stimulus materials is critical because this step directly impacts the effectiveness of emotion elicitation. Previous studies have employed various types of stimuli to evoke emotions, including music [57], pictures [58], facial expressions [15], and videos [28], [32]. Among all the available stimulus materials, videos have been found to be particularly effective because they provide both visual and auditory stimuli, which can reliably and efficiently elicit emotions.

During the preliminary stage, a pool of stimulus materials comprising video clips was prepared for eliciting six emotions, excluding surprise. A group of volunteers are requested to provide several videos (mainly 2 to 3) per emotion based on

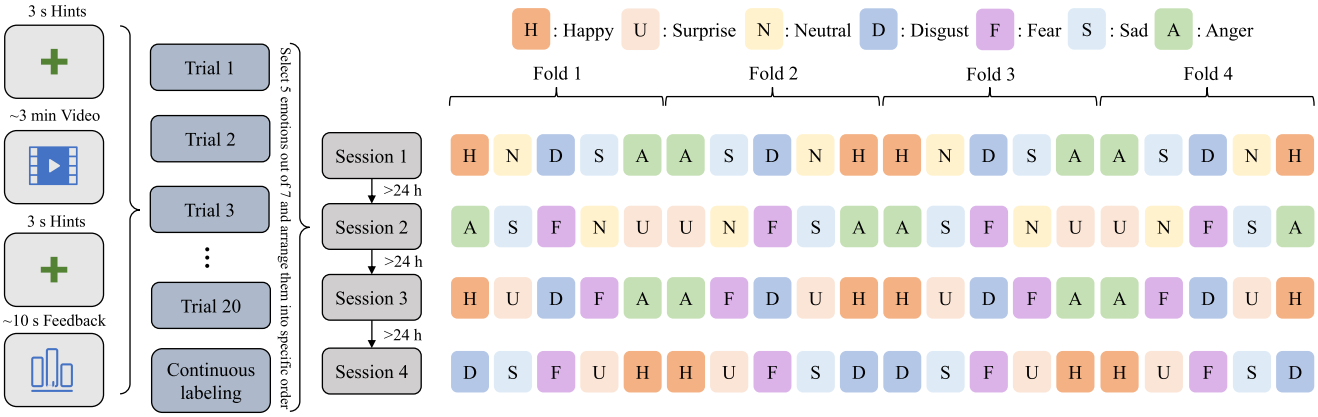


Fig. 1. The experimental design for video watching. Each row on the right represents a session that includes twenty trials. Different colors represent different emotions that are to be evoked in the videos. The leftmost side is the process of a single trial, which includes four parts: a starting hint, video watching, an ending hint, and feedback.

their subjective feelings, thus forming the stimulus pool. To select the most effective video clips for eliciting emotions, we employed a strategy involving the assessment of all the video clips by approximately 20 volunteers, who rated each clip on a scale ranging from 1 to 5. We selected the high-scoring clips for each emotion. To be consistent with prior SEED series datasets, we reused the videos in SEED series datasets for happy, sad, neutral, disgust, and fear emotions as they all underwent the same selection process mentioned above. For anger, we collected 44 video clips and finally chose the highest 12 ones. Considering that surprise could be either positive or negative, we primarily focus on neutral surprise to avoid confusion with happy, sad, or fear emotions. Magic videos were chosen for emotion elicitation, as magic shows have been demonstrated to be effective at eliciting surprise [38] and people tend to exhibit a neutral surprise. Consequently, twelve clips were selected for each emotion (except neutrality), with mean scores of 3 or higher. The neutral emotion comprised eight clips, resulting in a total of 80 video clips. Each video clip lasted for two to five minutes, and the total duration of all the clips was approximately 14,097.86 seconds. We elaborately separated the 80 clips into four parts, as shown in Fig. 1, and the subjects were required to complete the entire experiment in four sessions with intervals of 24 hours or more between sessions.

B. Subjects

Initially, 69 people from Shanghai Jiao Tong University signed up for the experiment through our recruitment questionnaire. To balance the sex ratio and select the most suitable subjects, only 24 subjects participated in our experiment. 4 of them dropped out of the experiment for individual reasons so their data were unavailable. Finally, only twenty subjects (10 males and 10 females) aged 19 to 26 years (mean: 22.5; STD: 1.80) participated in the experiments entirely with available data recorded. All participants were right-handed and had self-reported normal or corrected-to-normal vision and normal hearing at Shanghai Jiao Tong University. The participants were selected through the Eysenck Personality Questionnaire (EPQ), a widely used questionnaire developed by Eysenck et al. to assess an individual's personality traits [59]. Eysenck initially conceptualized personality as several biologically based independent temperament dimensions: E (extraversion/introversion), N (neuroticism/stability), P (psychoticism/socialization), and L (lie/social desirability). Previous research has demonstrated that individuals with extroverted characteristics perform better in terms of perceiving emotions during experiments than those without such characteristics [20], and people with high extraversion possess more empathy [60], [61]. Therefore, we ranked the volunteers according to their E values and selected those with high E values for the experiments. This approach was adopted to ensure that the participants possessed the desired characteristics necessary for accurately performing emotion recognition.

C. Protocol

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

All of the subjects underwent four experimental sessions. The procedure of the experiment is illustrated in Fig. 1. Twenty trials were included in each session; each trial consisted of two parts, where the first part involved watching videos and the second part involved self-assessment. Subjects scored their emotional intensity levels from 1 to 10 points. The self-assessment part was not time-limited and typically took approximately 10 seconds to complete.

For each session, only five out of seven emotions were elicited, which reduced the impact of subjects switching into too many emotional states. Prior to and following the presentation of each

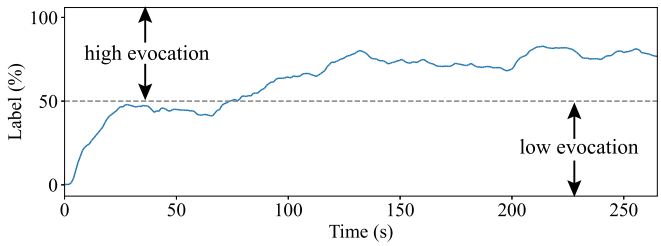


Fig. 2. The continuous labels annotated by the subjects.

video clip, a 3-second countdown was provided to alert participants to the imminent start or end of the video. The sequence of video clips presented in Fig. 1 was carefully arranged to avoid sudden emotional valence shifts, as human emotions tend to transform gradually.

Eighty video clips acquired from four sessions in total were divided into four folds. Each fold contained five clips from each session, and all the emotional videos were equal in number. At the conclusion of each session, the participants were instructed to review all twenty video clips, recall the emotional responses they experienced during the session and assign continuous labels to the entire session via a mouse wheel. The subjects were free to choose to review in real-time or speeded time. An example of continuous labels is presented in Fig. 2. The continuous labels ranged between 0% and 100%, where larger values corresponded to stronger elicited emotions.

D. Ethics Statement

This study was approved by the Scientific & Technical Ethics Committee of Shanghai Jiao Tong University. All subjects were informed of the experimental process before the first session and signed informed consent forms.

IV. Method

A. Multimodal Adaptive Emotion Transformer

The overall architecture of the Multimodal Adaptive Emotion Transformer is illustrated in Fig. 3. The training procedure has two phases. The model is first trained using both EEG and eye movement features to endow it with the ability to process multimodal inputs. Afterward, the backbone of the MAET is frozen, and emotional prompt tuning is introduced to tune only the emotional prompts and the classifier of a single modality. Once the MAET is trained, it can take either EEG or eye movement signals or both EEG and eye movement signals as its inputs. Given an input feature $x \in \mathbb{R}^d$, where d is the dimensionality of the feature, x is first passed to a multi-view embedding module to map the single feature to multiple tokens from different views. Then, the results are fed into an adaptive Transformer and a mixture Transformer, and the emotions are finally predicted by the classifiers.

1) Multi-View Embedding Module: The multi-view embedding module takes the input feature and transforms it into multiple embeddings, with the aim of encouraging 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 dimensionality of the 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 in an elementwise manner and then stacked over v embeddings, resulting in $E = (E_1, \ldots, E_v) \in \mathbb{R}^{v \times d_e}$.

The final output can be calculated as

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

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

Notably, the multi-view embedding module is optional for EEG signals because EEG features are naturally sequences formed by multiple channels or multiple frequency bands and can be applied directly by multi-head self-attention. However, we still adopt this module for the EEG data in this paper since we observe a performance boost when this module is included.

2) Adaptive Transformer and Mixture Transformer: The adaptive Transformer and mixture Transformer are flexible components that are inspired by the mixture-of-experts transformer [62]. These two modules are capable of covering arbitrary scenarios, such as inputs containing only EEGs, only eye movements, and both EEGs and eye movements, owing to the flexibility of 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 use it 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

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

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

The core components of the adaptive Transformer and mixture Transformer are the same, i.e., multi-head self-attention (MHSA) [42]. The embeddings $\tilde{E}$ are transformed to queries Q_i, keys K_i, and values V_i by three linear layers. The self-attention process can be calculated as

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

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

The adaptive Transformer introduces two modality experts to substitute for the standard feed-forward network (FFN),

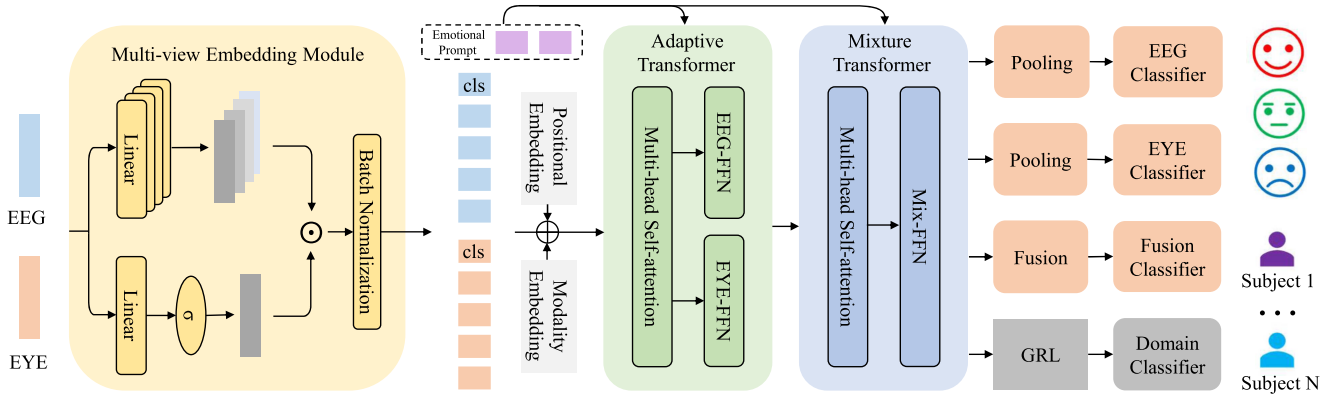


Fig. 3. The architecture of the MAET. The MAET is a general and flexible framework for EEG and eye movement signals that is composed of a multi-view embedding module, an adaptive Transformer block, a mixture Transformer block, and several classifiers.

i.e., EEG-FFN and EYE-FFN, and adaptively selects an expert to capture modality-specific information according to the input modality. For example, if the input is EEG-only (eye movements-only), we employ the EEG-FFN (EYE-FFN) expert 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 in parallel. The mixture Transformer follows the vanilla Transformer, where Mix-FFN is expected to capture more modality interactions. We stack L_a adaptive Transformer blocks and L_m mixture Transformer blocks.

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 approach to adaptively fuse the features derived from multiple modalities. We first calculate the attention weights μ^{eeg} and μ^{eye} by

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

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

$$H = \mu^{eeg} H_{cls}^{eeg} + \mu^{eye} H_{cls}^{eye}. \quad (7)$$

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 follows:

$$y^m = \mathcal{C}_f(\mathcal{F}(x^{eeg}, x^{eye})), \quad (8)$$

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

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

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

4) Emotional Prompt Tuning: We introduce emotional prompt tuning, which is inspired by the advent of prompt tuning [63], [64], to tune the model that has been trained on multimodal inputs to adapt to a single modality. The idea is quite straightforward. We prepend a small set of learnable embeddings $P_i \in \mathbb{R}^{p \times d_e}$, which are referred to as emotional prompts, to the feature embeddings in each Transformer layer. The emotional prompt tuning process can be formulated as

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

where TL_i denotes the ith Transformer layer and $\tilde{E}_i$ denotes the feature embeddings of the ith layer. $\tilde{E}_{i+1}$ is the output and the input of the $i+1$th Transformer layer. After all the adaptive and mixture Transformer layers, we adopt mean pooling over all the EEG or eye movement embeddings, followed by the use of classifier $\mathcal{C}_{eeg}$ for EEG signals or $\mathcal{C}_{eye}$ 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 the model learns to predict emotions using a single modality.

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 makes cross-subject emotion recognition challenging. To reduce the negative impacts of individual discrepancies, we exploit the adversarial domain generalization method to increase the robustness of the model [65]. 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 $\mathcal{C}_d$ that consists of two linear layers and a GELU [66] function between them. The domain classifier is trained jointly with the other components in the MAET to determine which domain the input belongs to. However, overconfident domain classifiers and domain label noise can lead to instability during the domain-adversarial training process. To overcome this challenge, we adopt environment label smoothing (ELS) [67], which encourages the domain classifier to output soft probabilities.

For a domain label $d \in [0, 1]^K$, we transform it to $\hat{d}$ as follows:

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

where i ranges from 1 to K and $\sum_{i=1}^{K} \hat{d}(i) = 1$. γ is the tradeoff that controls the convergence of the algorithm and minimizes the adversarial divergence. We follow the annealing strategy [67] that gradually decreases γ during the training process as $\gamma = 1 - \frac{K-1}{K}\frac{t}{T}$, where t is the current training step and T is the total number of steps.

Therefore, the loss of the domain classifier is

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

To confuse the domain classifier so that the feature extractor can learn domain-invariant representations, we introduce a gradient reversal layer (GRL) [65], which can be ignored during forward propagation and reverses the gradient that passes backward from $\mathcal{C}_d$ to $\mathcal{F}$. Consequently, the total loss for the EEG-based cross-subject emotion recognition task 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 during the early stages of the training procedure.

B. Feature Extraction

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 [68]. We then apply a bandpass filter with cutoff frequencies of 0.1 Hz and 70 Hz to remove low-frequency noise. Additionally, a notch filter with a cutoff frequency of 50 Hz is applied to prevent powerline interference. To reduce the computational complexity of our method, 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 handcrafted feature for emotion recognition, as it has a balanced ability to discriminate between EEG patterns with low- and high-frequency energy [69]. We use a 256-point Short-Time Fourier Transform (STFT) with a non-overlapping 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, and gamma: 31–49 Hz), which are defined as

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

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

Based on the assumption that emotional states are defined in a continuous space and that emotional states change gradually over time, we exploit the linear dynamic system (LDS) approach to filter out components that are not associated with emotional states [70].

2) Eye Movement Features: Various eye movement parameters, such as pupil diameters, fixation details, saccade details, and gaze point details, can be captured by eye gaze trackers. Among these parameters, pupil diameters have been demonstrated to play a critical role in emotion recognition [71]. Nonetheless, pupil diameters are highly sensitive to environmental luminance. We first employ linear interpolation to replace the pupil diameter samples that are missing due to eye blinking.

Based on the observation that the responses of subjects to the same video in a controlled lighting environment have similar patterns, principal component analysis (PCA) is used to eliminate the effect of luminance on the 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 the pupil diameter data obtained for the same video clip from all subjects. After that, the residual part contains only the emotion-associated pupil responses in addition to noise.

The DE features are subsequently computed for the left and right pupil diameters using the STFT in four frequency bands (0–0.2 Hz, 0.2–0.4 Hz, 0.4–0.6 Hz, and 0.6–1 Hz) with a non-overlapping Hanning window of 4 seconds. In addition to the DE features, the mean and the standard deviation of the pupil diameters are also calculated. In addition to the pupil diameters, twenty-one other features are extracted, as shown in Table II [29]. Consequently, the total number of features obtained from the eye movement signals is 33.

TABLE II
EYE MOVEMENT FEATURES

Eye movement parameters	Dimensions	Extracted features
Pupil diameters (left and right)	1-12	DE in four bands (0-0.2 Hz, 0.2-0.4 Hz, 0.4-0.6 Hz, and 0.6-1 Hz), mean, standard deviation
Fixation duration (ms)	13-14	Mean, standard deviation
Dispersion (X and Y)	15-18	Mean, standard deviation
Saccade duration (ms)	19-20	Mean, standard deviation
Saccade amplitude (°)	21-22	Mean, standard deviation
Blink duration (ms)	23-24	Mean, standard deviation
Event statistics	25-33	Blink frequency, fixation frequency, maximum fixation duration, maximum fixation dispersion, total fixation dispersion, saccade frequency, average saccade duration, average saccade amplitude, average saccade latency

TABLE III
THE ACCURACIES AND F1 SCORES (AVG./STD., %) ACHIEVED BY DIFFERENT METHODS USING EEG OR EYE MOVEMENT SIGNALS

	Method	EEG signals						Eye movements
		delta band	theta band	alpha band	beta band	gamma band	all bands	
Accuracy	KNN [72]	28.21/5.76	28.47/7.02	31.92/6.51	34.83/5.77	37.20/6.06	36.43/**5.38**	36.01/**6.30**
	HCNN [36]	42.33/6.36	42.63/5.28	43.62/**4.82**	47.79/6.77	48.18/6.10	52.42/6.47	-
	RGNN [37]	42.31/5.14	41.73/5.53	43.93/4.98	44.68/**5.51**	45.49/**4.73**	48.50/6.83	-
	Transformer [42]	46.87/**4.10**	47.01/**4.23**	48.78/5.39	53.10/6.34	53.96/6.60	56.04/7.82	-
	GCNCA [38]	**49.97**/4.94	**50.46**/4.29	**53.26**/5.45	**58.36**/6.69	**59.50**/5.77	58.04/7.78	-
	MAET	-	-	-	-	-	**58.11**/8.78	**50.31**/7.14
F1 score	KNN [72]	26.00/5.98	26.12/7.25	29.67/6.64	33.22/**5.59**	34.98/5.62	34.08/**5.79**	34.74/**6.33**
	HCNN [36]	37.85/7.78	38.30/5.72	39.66/**5.23**	44.05/7.55	44.33/6.68	49.02/6.80	-
	RGNN [37]	37.85/5.56	36.13/5.79	39.18/5.43	40.68/6.61	41.77/**5.50**	45.32/7.20	-
	Transformer [42]	43.20/**4.26**	43.85/**4.74**	45.48/**5.23**	49.81/6.70	51.48/6.96	53.35/8.30	-
	GCNCA [38]	**47.61**/5.14	**47.92**/5.01	**50.99**/5.71	**55.97**/7.12	**57.54**/5.90	**55.48**/8.30	-
	MAET	-	-	-	-	-	54.98/9.45	**47.10**/7.84

V. EXPERIMENTAL RESULTS

A. Implementation Details

Regarding the hyperparameters of the MAET, the numbers of adaptive Transformer blocks L_a and mixture Transformer blocks L_m are 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 4in the MHSA module. The embedding dimensionality d_e is tuned within $\{32, 64\}$. The batch size is 64in the subject-dependent experiments and 256in the cross-subject experiments. We use AdamW [73] as the optimizer, with its learning rate tuned within $\{0.00003, 0.0001, 0.0003\}$. Moreover, we tune the weight decay within $\{0.0001, 0.01, 0.1\}$. The prompt length p is 1 or 2. Note that domain adversarial training is employed only under cross-subject conditions. Emotional prompt tuning is only employed in Section V-B. Otherwise, the MAET is directly trained using only the EEG features with the cross-entropy loss function.

B. Unimodal and Multimodal Emotion Recognition

To evaluate the efficacy of EEG and eye movement signals in terms of identifying the seven target emotions, we construct a subject-dependent model for each subject. Specifically, we merge the data acquired from one subject during all four sessions to train the models and then partition the data into four folds for carrying out a four-fold cross-validation process, as illustrated in Fig. 1. The overall performance of our proposed methods is determined based on the average fourfold cross-validation results. Notably, the input EEG and eye movement features are transformed by z score normalization. The experimental results are shown in Tables III and IV.

1) Classification Performance of EEG Signals: With the extraction of differential entropy from five individual frequency bands and the total EEG band (delta, theta, alpha, beta, and gamma), we further investigate the critical bands for the seven emotions by conducting classification tasks on each band as well as on the total band. The classification performances of six existing classifiers, K-nearest neighbors (KNN) [72] (K is set to 1), a hierarchical convolutional neural network (HCNN) [36], a regularized graph neural network (RGNN) [37], a Transformer [42], and a graph convolutional network with channel attention (GCNCA) [38] are systematically compared with a newly developed neural network called the multimodal adaptive emotion transformer (MAET) in this paper. Note that the EmotionDL algorithm proposed in the RGNN is not implemented in this paper. All methods are implemented strictly under the same conditions and are fairly compared with each other. Table III shows the average accuracies and F1 scores produced by each method.

TABLE IV
THE ACCURACIES AND F1 SCORES (AVG./STD., %) PRODUCED BY DIFFERENT METHOD USING MULTIMODAL SIGNALS

Method	Accuracy		F1 score	
	Avg.	Std.	Avg.	Std.
KNN [72]	40.44	**6.30**	37.94	**6.54**
BDAE [74]	61.55	8.74	59.11	8.87
ETF [41]	65.30	8.55	63.13	8.88
VigilanceNet [75]	62.93	7.12	60.46	7.81
MAET	**71.28**	7.74	**69.16**	8.35

Differences between bands with different frequencies: Notably, high-frequency bands, namely, the alpha, beta, and gamma bands, exhibit superior discrimination capabilities in comparison to those of the low-frequency bands (the delta and theta bands) for identifying the seven emotions. Moreover, the band frequency is positively correlated with performance across the five bands. Additionally, the gamma band outperforms the other isolated frequency bands with all classifiers, while the total band yields the best performance among the single bands in most cases, suggesting complementary properties among the distinct frequency bands. Notably, the superior discrimination ability of the gamma band is a new finding that contrasts with previous research [20] involving fewer emotions, where the beta band was found to be the most effective band. This disparity may arise from the fact that the gamma band has more pronounced discriminative properties, especially for emotions that have not been extensively studied before. This discovery suggests that more emotion-associated information might be contained in the gamma band or the band with the highest frequency, which highlights the need for future research to pay more attention to the gamma band because of the precision, complexity, and variety of emotions.

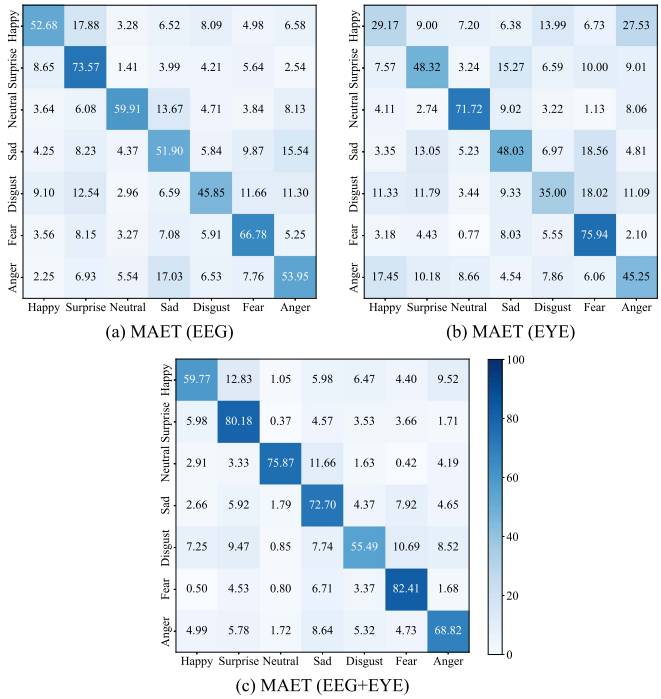


Fig. 4. Confusion matrices produced by the MAET using EEG signals, eye movements or both. The horizontal axis represents the predicted labels, and the vertical axis represents the true labels.

Differences between different models: In the task of classifying the seven emotions, deep learning methods with attention mechanisms are absolutely stronger than traditional models such as KNN. Utilizing low-frequency bands, GCNCA achieves the best classification accuracy compared to those of the other models. In contrast, when using high-frequency bands, the MAET attains the most accurate discrimination results. Specifically, the highest prediction accuracy of 58.11% is acquired by the MAET while utilizing the total band, highlighting the effectiveness of the MAET.

Fig. 4(a) depicts the confusion matrix produced by the MAET using only EEG signals. The surprise and fear emotions are more accurately distinguished by the MAET than the other emotions. In addition, the happy emotion is prone to be misclassified as surprise, while the neutral emotion is more likely to be confused with sadness. Furthermore, compared to the other emotions, the sad and angry emotions are more likely to be misclassified as each other during the classification process implemented based on EEG signals, which indicates the similarity between the neural patterns of the sad and angry emotions.

2) Classification Performance of Eye Movements: For the eye movement experiment, the 33-dimensional features extracted as described in Section IV-B2 are employed for classification purposes. We compare our proposed MAET with KNN since the other baseline methods are unable to handle eye movement features. The results are shown in Table III. Notably, our MAET model yields the highest prediction accuracy (50.31%), which is far greater than the 36.01% achieved by KNN. Additionally, the highest F1 score of 47.10% is achieved by our MAET model.

Fig. 4(b) presents the confusion matrix produced by the MAET when using eye movement signals alone. It is evident from the table that eye movement signals exhibit remarkable performance in terms of distinguishing between neutral and fearful emotions. Nonetheless, isolated eye movement signals have relatively poor performance with respect to classifying the happy and disgust emotions, whose accuracies are lower than 40%. As shown in Fig. 4(a) and (b), EEG signals attain better results in terms of discriminating happiness, surprise, disgust, and anger, while eye movement signals acquire more accurate results when classifying neutrality, sadness, and fear. Notably, some similar eye movement patterns are observed between the happy and angry emotions because 27.53% of the happy emotions are recognized as anger.

3) Classification Performance of Multimodal Signals: Table IV displays the experimental results obtained by different models using both EEG signals and eye movements. A systematic comparison is conducted between KNN, the Bimodal Deep Autoencoder (BDAE) [74], Emotion Transformer Fusion (ETF) [41], VigilanceNet [75], and the MAET. For KNN, the EEG features and eye movement features are directly concatenated into 343-dimensional feature vectors. The MAET outperforms the other classifiers, with the best prediction accuracy of 71.28% and the best F1 score of 69.16%, which illustrates the effectiveness of our model. Moreover, ETF and VigilanceNet reach the second- and third-highest accuracies of 65.30% and 62.93%, respectively, as both of these methods utilize attention mechanisms, as does our MAET model, demonstrating that an attention mechanism is a significant component for classifying emotions using multimodal signals.

The confusion matrix produced by the MAET using multimodal signals is shown in Fig. 4(c), which shows the details of its ability to discriminate among the seven emotions. The MAET achieves outstanding accuracy in terms of classifying the surprise, neutral, fear, and angry emotions. Among all seven emotions, the fear emotion yields the highest discrimination accuracy of 82.41%. Most emotions can be classified more accurately when multimodal signals are used than when EEG or eye movement signals are used individually. The accuracy achieved when discriminating happy emotions using multimodal signals reaches 59.77% compared to 52.68% when only EEG signals are utilized, which is an increase of 7%. In addition, the accuracies achieved when discriminating sadness, fear, and anger increase by 21.70%, 15.63%, and 14.87%, respectively. These results demonstrate that jointly utilizing EEG and eye movement signals can significantly improve the classification performance of the model, which indicates the complementary properties of EEG and eye movement signals in terms of recognizing emotions.

C. Cross-Subject EEG-Based Emotion Recognition

One of the essential questions in EEG-based emotion recognition is whether this approach is reliable and robust when recognizing the emotions of a new subject whose physiological signals have never been recorded and fed into classifiers. Many factors, such as gender, age, cultural background and the specific emotion states elicited by stimulus materials, likely influence

TABLE V
THE PERFORMANCE ACHIEVED BY DIFFERENT METHODS IN CROSS-SUBJECT EXPERIMENTS

Method	Accuracy		F1 score	
	Avg.	Std.	Avg.	Std.
KNN [72]	20.85	4.56	20.23	4.49
HCNN [36]	39.88	4.94	38.18	5.06
RGNN [37]	37.49	5.44	34.52	4.83
Transformer [42]	40.36	5.22	37.76	5.54
GCNCA [38]	38.68	**3.94**	37.25	**3.65**
CLISA [76]	38.27	5.23	34.02	5.34
MAET (w/o AT)	40.69	5.50	38.47	6.09
MAET	**40.90**	5.52	**38.85**	6.07

the classification accuracy differences observed among different subjects during the experiments. To further investigate the performance of our MAET model when facing the above problems, we compare our MAET model with other classification methods. The strategy we adopt for measuring the achieved cross-subject performance is leave-one-subject-out (LOSO) cross-validation. For each subject, a model is trained with the data from the other 19 subjects used as the training set and the particular data from the target subject used as the test set. Afterward, all 20 test results are consolidated to calculate the average accuracy. Apart from the default baselines, we consider a contrastive learning-based approach tailored for learning subject-invariant EEG representations called CLISA [76].

The results are depicted in Table V. It can be seen from the table that the deep methods are more reliable and robust than the traditional methods such as KNN in the cross-subject experiment. Due to the variability between distinct subjects, the performances of all methods are worse than those achieved under subject-dependent conditions, and the performance degradation is nearly 20%. The lowest standard deviation is achieved by GCNCA, with accuracy and F1 score values of 3.94% and 3.65%, respectively. Our MAET model achieves the highest accuracy of 40.90% and an F1 score of 38.85%, demonstrating the robustness of the MAET for cross-subject emotion recognition. Notably, the MAET without adversarial training (AT) attains the second-highest accuracy of 40.69%, which implies that adversarial training is helpful for addressing cross-subject situations to some extent.

D. Neural Signatures and Stable Patterns

To further explore the particularity of the neural signatures associated with the seven emotions, we project the DE features to the scalp to determine stable neural patterns. The DE features are first transformed by z score normalization for each subject and then averaged over all subjects. Fig. 5 shows the DE topographic maps produced for the seven emotions in five distinct frequency bands.

The results show that the happy emotions in the beta band and gamma band exhibit greater activation in the lateral temporal areas than do all other emotions; moreover, the energy in the prefrontal area is significantly lower for happy individuals than for all negative emotions (sadness, anger, disgust, and fear). For the surprise emotion, the most distinguishable feature is that the energy of surprise is particularly low in all frequency bands, which illustrates why its prediction accuracy of 73.57% is the highest among those of all emotions when using EEG signals alone. The neural pattern of the neutral emotion involves strong alpha responses in the parietal and occipital areas, along with a small area with low energy in the vertex of the cerebral cortex. Existing studies [77], [78] have shown that the EEG alpha band response is correlated with attention, and less attention results in a high alpha band response. When presented with neutral videos, subjects are prone to relaxing and paying less attentional, which leads to high alpha responses. This finding also demonstrates that people with surprise emotions had significantly poorer alpha responses since they definitely concentrated on the magic videos.

For the negative emotions, including sadness, anger, disgust, and fear, we can summarize the following neural patterns. The energy contained in the lateral temporal areas is low, while the prefrontal area obtains high energy in the gamma band for sadness. Regarding anger, in the gamma band, the DE features are quite low in the temporal and frontal areas. In the theta, alpha, and beta bands, the neural patterns exhibit moderate activation in the occipital areas, and a small area with low energy is observed in the vertex of the cerebral cortex, which is the same position as that of the neutral emotion. Notably, for the disgust emotion, a small area with higher energy than other areas is observed in most frequency bands except the gamma band, while the occipital areas exhibits less activation than the other areas. The most detectable neural pattern of fear is that the frontal cortex has strong activation in all frequency bands, especially in low-frequency bands such as the delta and theta bands. Moreover, the parietal areas of fear exhibit less activation, while the temporal and occipital areas have relatively high activation levels in all bands. The findings related to these emotions are consistent with previous fMRI findings [33].

From the DE topographic maps, we observe that the specific neural patterns of the seven emotions exist in the high-frequency beta and gamma bands. However, more obstacles are encountered when classifying emotions in the low-frequency bands because some emotions, such as happiness, sadness, anger, and disgust, exhibit moderate and vague activation levels, which explains the poor prediction accuracy achieved by using low-frequency bands.

E. Statistics of Eye Movement Signals

We analyze four typical eye movement features, namely, the pupil diameter, fixation duration, saccade duration, and dispersion of X, for each emotion. The distributions of these features are represented using violin plots, as depicted in Fig. 6. Compared to other eye movement features, the most distinguishable feature among the four is the pupil diameter, while the fixation duration, saccade duration, and dispersion of X have relatively low discriminability. Among all the emotions, the neutral emotion possesses the smallest pupil diameter, which is in line with the findings of previous studies [79], [80] concluding that attentional lapses are more likely to occur when the pupil diameter is small. This result is consistent with the analysis in Section V-D stating that people in neutral emotional states tend

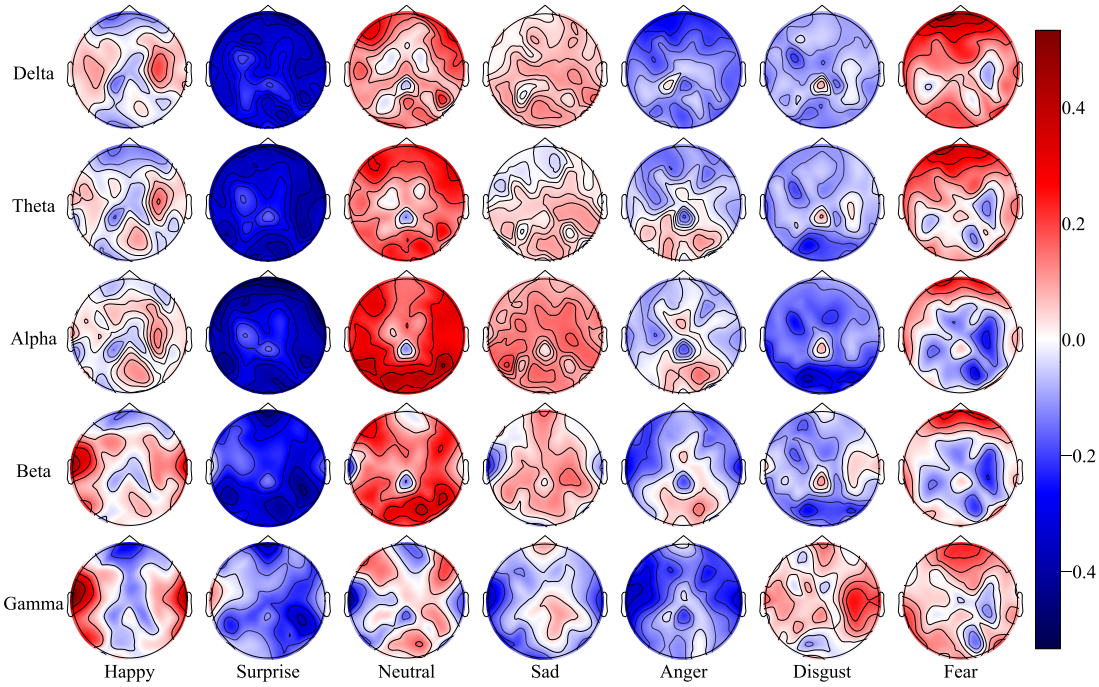


Fig. 5. The DE topographic maps produced for seven emotions in five frequency bands. The DE features are first normalized for each subject and then averaged over all subjects.

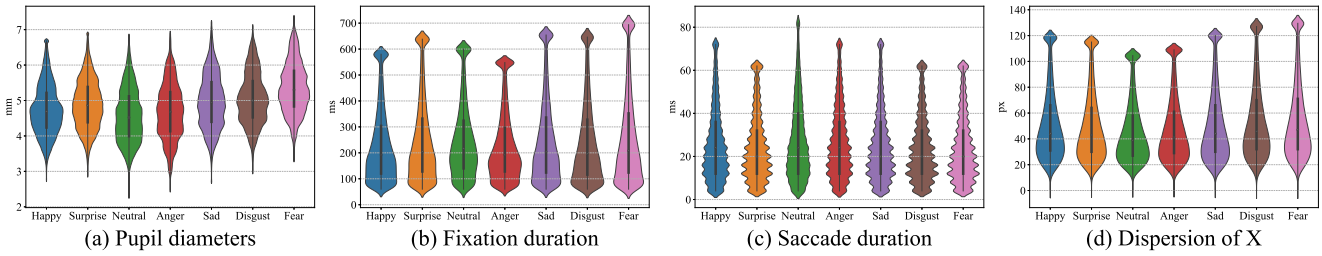


Fig. 6. Violin plots of several eye movement statistics. The white points indicate the median values, and the bold black lines indicate the interquartile ranges.

to be less attentional. Conversely, fear exhibits the greatest pupil diameter, while no significant difference is observed between the pupil diameters of the happy and angry emotions. Moreover, the surprise, sad, and disgust emotions have similar pupil diameters. The fixation duration for the neutral emotion is the longest among all emotions, whereas people with the happy, angry, and disgust emotions fixate more momentarily. For the saccade duration, the happy, neutral, and angry emotions have similar durations, which are slightly longer than the durations of the other four emotions. The largest dispersion of X exists for people with fear, and the neutral emotion has the smallest dispersion of X. These findings are consistent with those of a previous study that considered five emotions [50].

F. Analysis of Continuous Labels

The intensity score associated with an emotion is a crucial indicator of its elicitation level and the quality of the collected physiological data. Several previous studies [34] have evaluated emotions through the use of continuous labels, which can sensitively measure affective arousal. In our study, we employed a continuous intensity level rating scale, wherein participants were asked to score their f elicitation levels by means of a mouse wheel while reviewing all video clips at the end of each session. For each particular video clip, the intensity scores of the 20 subjects are averaged, and the result is depicted in Fig. 7; the order of the 80 videos is equivalent to that used in Fig. 1. Due to the ambiguous intensity score definition for neutrality, we exclude the neutral emotion from our analysis in this experiment. The middle lines represent the average scores, the color depths of which represent the numbers of subjects predicted by the MAET to be highly induced. Moreover, the colored areas depict the standard deviations of the intensity scores. To better observe the details of the continuous labels, the range of the Y-axis is scaled to a more appropriate size for each video clip.

1) Effectiveness of Filtering High-Induced Data: To further investigate the relevance of the differences between classification performance and intensity, we conduct a comparison

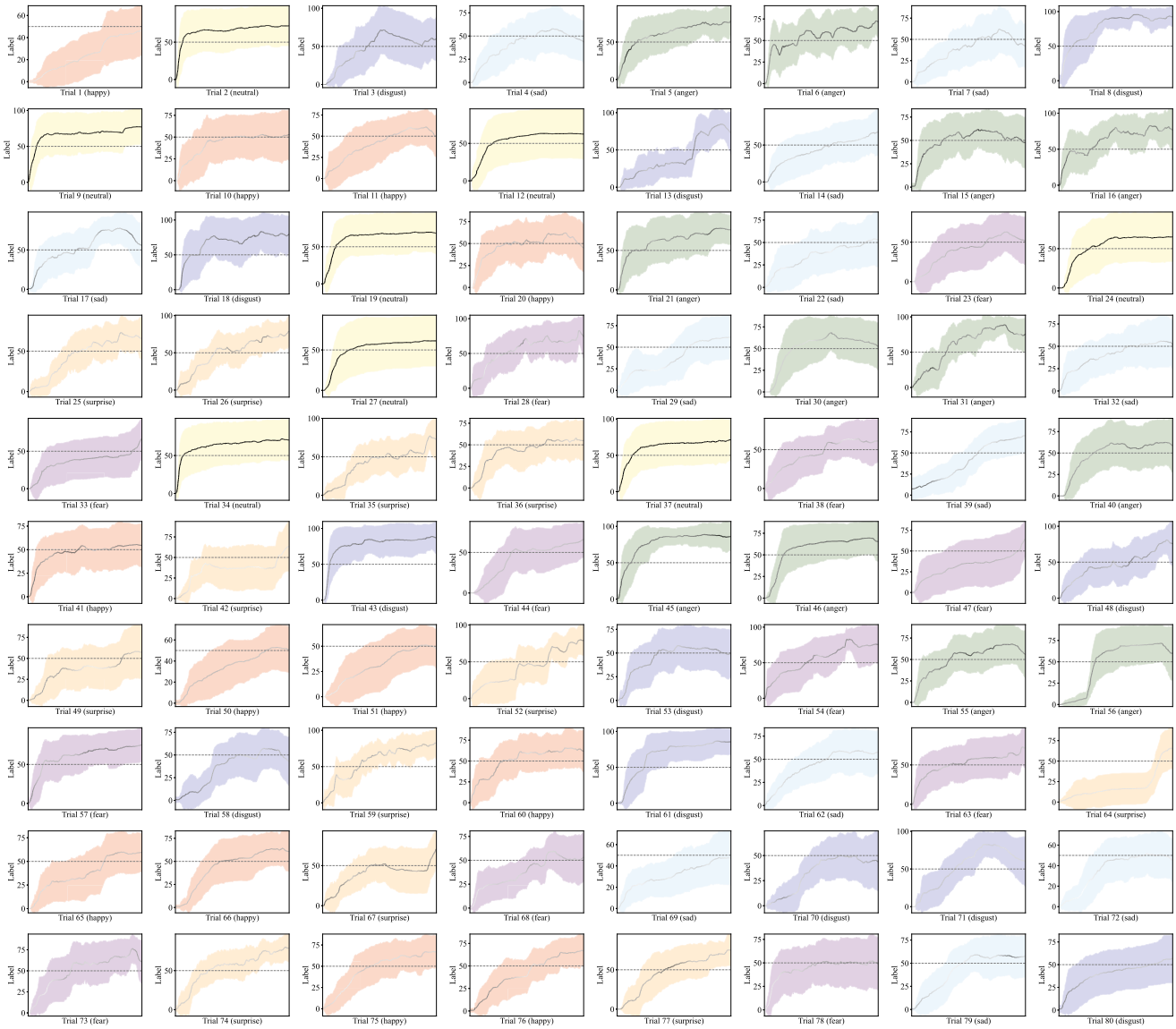


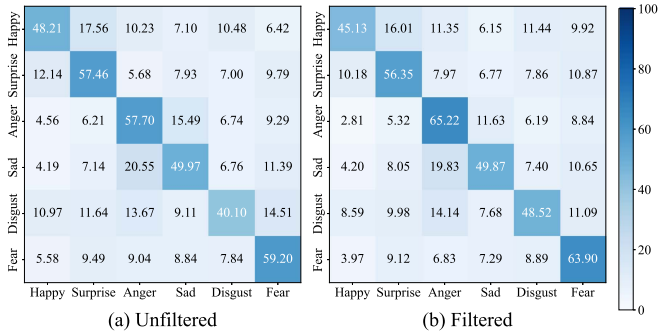
Fig. 7. Continuous labels annotated by the participants. The middle lines indicate the average scores, and the darker the lines are, the more subjects are predicted by the MAET to be highly induced. The colored areas represent the scores located within the standard deviations.

among the classification performances achieved by each method under unfiltered and filtered situations, and low-induced data are filtered out in the filtered situation. The criterion for judging whether EEG signals are strongly induced is that the score of the corresponding video clip must be greater than a threshold of 50%, which indicates the part above the horizontal dashed line shown in Fig. 7. We adopt a smoothing algorithm called LDS for improving the classification performance; this approach can smooth the DE features in a particular video clip, as described in Section IV-B1. For the purpose of discriminating between the quality of high-induced and low-induced data, LDS is not applied in this experiment. The experimental results are displayed in Table VI. It is evident that rapid accuracy and F1 score declines occur in the unfiltered case compared to the results in Table IV. From Table VI, we can see that the classification accuracy and F1 score increase considerably merely by using the filtered data for all methods.

TABLE VI
THE ACCURACIES AND F1 SCORES YIELDED BY DIFFERENT METHODS WITH AND WITHOUT THE FILTERED HIGH-FREQUENCY DATA USING CONTINUOUS LABELS

Method	Unfiltered		Filtered	
	Accuracy	F1 score	Accuracy	F1 score
KNN [72]	31.39	29.92	33.98	32.86
HCNN [36]	46.45	44.29	50.24	45.78
RGNN [37]	43.61	42.51	50.70	49.55
Transformer [42]	49.71	48.71	56.84	56.19
GCNCA [38]	52.74	52.21	58.15	**58.16**
MAET	**52.86**	**52.26**	**58.24**	58.08

Fig. 8 depicts the confusion matrices produced by the MAET in the unfiltered and filtered situations. By filtering the high-induced data using continuous labels, the prediction accuracies achieved for anger, disgust, and fear increase by 7.52%, 8.42%,



(a) Unfiltered (b) Filtered

Fig. 8. Confusion matrices produced by the MAET with and without the filtered high-induced data using continuous labels.

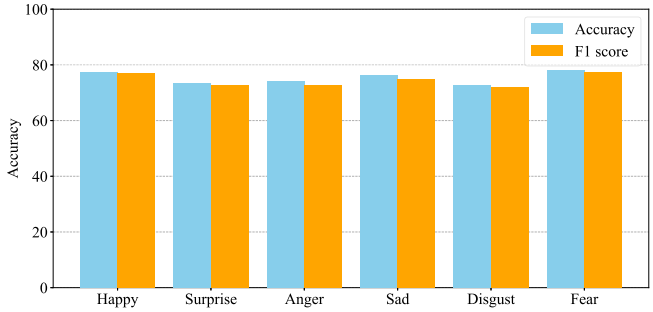


Fig. 9. The accuracies and F1 scores produced for seven emotions in high-/low-induced emotion classification scenarios.

and 4.7%, respectively, which underscores the importance of performing filtering for discriminating among these three emotions. However, slight classification accuracy decreases are observed for the happy, surprise, and sad emotions. Fig. 7 shows that for the happy, surprise, and sad emotions, it takes longer for subjects to be evoked by stimulus materials, which results in a lack of physiological data after filtering. As deep models require large amounts of data, the use of inadequate data after filtering may account for the decrease in the ability of the model to discriminate among these three emotions. This observation further demonstrates the effectiveness of filtering for happy, surprise, and sad emotions. These findings suggest that filtering high-induced data is important for classifying easily evoked emotions.

2) Classification of High-/Low-Induced Emotions: The significance of filtering high-frequency data using continuous labels has been demonstrated in the previous section, highlighting the potential of this approach to provide enhanced classification performance. Nonetheless, manually acquiring continuous labels from subjects is time-consuming and prone to error. Notably, compared to the specific intensity score, we are more concerned about whether particular data are highly induced. Therefore, it is not necessary to carry out regression tasks to predict the specific intensity score of each dataset. The issue is simplified to a binary classification task of determining whether a given data point is highly induced. This classification experiment was performed using the MAET by every subject for six emotions each, except for the neutral emotion. We present the emotion-specific classification accuracies achieved by the MAET in Fig. 9, with no significant differences observed among the six emotions. The mean accuracy for all emotions is 75.33%, while the mean F1 score is 74.47%. Many factors may affect the accuracy and precision of the emotion states recalled by the subjects after the whole session, the distinct neural patterns between the six emotions, and the variety of emotions exhibited by different subjects. Nonetheless, the use of deep learning algorithms to discriminate high-induced or low-induced data is efficient and inexpensive, which suggests that future work should pay more attention to constructing better classifiers.

3) Regression of Emotional Intensity: In addition to the binary classification for emotional intensity, we also perform the regression task. Mean square error (MSE) loss is employed to optimize the models. Root mean square errors (RMSE) and Pearson's Correlation Coefficient (PCC) are utilized as evaluation metrics. Note that we use a linear transform (a linear layer followed by a ReLU activation) instead of the multi-view embedding module for EEG inputs in this experiment since the simple linear transformation achieves a better performance. Table VII exhibits the regression results of different methods. The RMSE of different methods are relatively close to each other (around 0.24), while the best PCC is slightly over 0.3. These results indicate the feasibility of directly regressing the emotional intensity for each emotion and it still has room for improvement in the future.

TABLE VII
THE ROOT MEAN SQUARE ERRORS AND PEARSON'S CORRELATION COEFFICIENT (AVG./STD.) OF EMOTIONAL INTENSITY REGRESSION

Method	RMSE↓		PCC	
	Avg.	Std.	Avg.	Std.
SVR	0.2560	0.0343	0.1561	0.0946
HCNN [36]	0.2404	0.0329	0.2397	0.0705
RGNN [37]	0.2527	0.0353	0.1887	**0.0536**
Transformer [42]	0.2360	0.0317	**0.3167**	0.0563
GCNCA [38]	0.2434	0.0332	0.2692	0.0544
MAET	**0.2358**	**0.0309**	0.3015	0.0698

G. Discussion

Our findings revealed distinct neural patterns associated with each of the seven emotions in the EEG data. For instance, the happy emotion showed higher activation levels in the temporal areas in the beta and gamma bands. Surprisingly, the surprise emotion produced very low DE features across all bands, suggesting a unique neural signature. The sad and angry emotions exhibited lower DE features in the gamma band in the temporal areas compared to other emotions. Additionally, the occipital area for the disgust emotion had the lowest DE features among all emotions, and the frontal area for the fear emotion showed high activation. These findings support the existence of specific neural signatures for different emotions, which is critical for developing accurate emotion recognition systems. Among the four eye movement features analyzed, the pupil diameter proved to be the most distinctive. Fear elicited the largest pupil diameter,

whereas the neutral emotion had the smallest. The fixation duration was longest for the neutral emotion, and shorter for emotions like happiness, anger, and disgust. The saccade duration was similar for happiness, neutrality, and anger, but shorter for the other emotions. The dispersion of X was greatest for fear and smallest for neutrality, reflecting attentional differences between emotional states.

The confusion matrices produced by the MAET using EEG signals alone showed that the surprise and fear emotions were more accurately distinguished than other emotions. However, the happy emotion was often misclassified as surprise, and the neutral emotion tended to be confused with sadness. Furthermore, the sad and angry emotions were frequently misclassified as each other, indicating a certain level of similarity in their neural patterns.

We also explored the impact of filtering data according to continuous labels, which represent the intensity of emotions. The results indicated that filtering high-intensity data led to a significant increase in classification accuracy, particularly for easily evoked emotions. This finding underscores the importance of focusing on emotionally intense samples for improving emotion recognition performance.

VI. Conclusion

In this study, we developed a novel multimodal emotion dataset named SEED-VII comprising seven emotions (happiness, sadness, fear, disgust, surprise, anger, and neutrality) with EEG and eye movement signals. An important feature of SEED-VII is that it includes continuous labels that indicate the affective intensity levels that subjects experienced during watching videos.

We proposed a novel multimodal Transformer model named the MAET, which is capable of flexibly addressing unimodal and multimodal inputs. The performances of different existing methods were systematically evaluated in unimodal and multimodal cases. Furthermore, we conducted a cross-subject experiment using LOSO cross-validation to evaluate the performance of each method.

The experimental results indicated that neural signatures and stable EEG patterns existed for the seven emotions, validating the feasibility of cross-subject research. We found that the happy emotion exhibited greater activation levels in temporal areas in the beta and gamma bands, the surprise emotion yielded very low DE features in all bands, the neutral emotion produced strong alpha responses, the temporal areas of the sad and angry emotions in the gamma band were lower than those of other emotions, the occipital area of the disgust emotion was the lowest among those of all emotions, and the frontal area of the fear emotion displayed high activation.

Moreover, a comparison between unfiltered and filtered situations was carried out to explore the effect of filtering data according to continuous labels. The experiments indicated that a considerable increase in accuracy was achieved with the filtered data. We discovered that filtering high-induced data is important for classifying easily evoked emotions.

References

[1] P. Salovey and J. D. Mayer, "Emotional intelligence," *Imagination Cogn. Pers.*, vol. 9, no. 3, pp. 185–211, 1990.

[2] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[3] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: Principles, state-of-the-art, and challenges," *Brain-Comput. Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

[4] D. Wu, B.-L. Lu, B. Hu, and Z. Zeng, "Affective brain–computer interfaces (aBCIs): A tutorial," in *Proc. IEEE*, vol. 111, no. 10, pp. 1314–1332, Oct. 2023.

[5] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

[6] D. Watts, R. F. Pulice, J. Reilly, A. R. Brunoni, F. Kapczinski, and I. C. Passos, "Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis," *Transl. Psychiatry*, vol. 12, no. 1, 2022, Art. no. 332.

[7] S. K. Loo, J. J. McGough, J. T. McCracken, and S. L. Smalley, "Parsing heterogeneity in attention-deficit hyperactivity disorder using EEG-based subgroups," *J. Child Psychol. Psychiatry*, vol. 59, no. 3, pp. 223–231, 2018.

[8] C. Imperatori et al., "Default mode network alterations in individuals with high-trait-anxiety: An EEG functional connectivity study," *J. Affect. Disord.*, vol. 246, pp. 611–618, 2019.

[9] T. L. Burleigh, M. D. Griffiths, A. Sumich, G. Y. Wang, and D. J. Kuss, "Gaming disorder and internet addiction: A systematic review of resting-state EEG studies," *Addictive Behaviors*, vol. 107, 2020, Art. no. 106429.

[10] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cogn. Emotion*, vol. 23, no. 2, pp. 209–237, 2009.

[11] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affective Comput.*, vol. 10, no. 3, pp. 374–393, Third Quarter, 2019.

[12] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, 2020.

[13] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[14] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 801–804.

[15] F. Z. Canal et al., "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Inf. Sci.*, vol. 582, pp. 593–617, 2022.

[16] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, and S. D. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network," *Neural Netw.*, vol. 18, no. 4, pp. 423–435, 2005.

[17] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights," *Procedia Comput. Sci.*, vol. 175, pp. 689–694, 2020.

[18] T. Sapiński, D. Kamińska, A. Pelikant, and G. Anbarjafari, "Emotion recognition from skeletal movements," *Entropy*, vol. 21, no. 7, pp. 1–16, 2019.

[19] Z. Shen, J. Cheng, X. Hu, and Q. Dong, "Emotion recognition based on multi-view body gestures," in *Proc. 2019 IEEE Int. Conf. Image Process.*, 2019, pp. 3317–3321.

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

[21] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[22] S. A. Mithbavkar and M. S. Shah, "Analysis of EMG based emotion recognition for multiple people and emotions," in *Proc. IEEE 3rd Eurasia Conf. Biomed. Eng. Healthcare Sustainability*, 2021, pp. 1–4.

[23] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Emotion recognition from facial EMG signals using higher order statistics and principal component analysis," *J. Chin. Inst. Engineers*, vol. 37, no. 3, pp. 385–394, 2014.

[24] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ECG-based emotion recognition in music listening," *IEEE Trans. Affective Comput.*, vol. 11, no. 1, pp. 85–99, First Quarter, 2020.

[25] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz, "Electrocardiogram-based emotion recognition systems and their applications in healthcare–A review," *Sensors*, vol. 21, no. 15, pp. 1–37, 2021.
[26] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multimodal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.
[27] F. L. da Silva, "EEG and MEG: Relevance to neuroscience," *Neuron*, vol. 80, no. 5, pp. 1112–1128, 2013.
[28] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, First Quarter, 2012.
[29] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.
[30] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.
[31] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
[32] S. Koelstra et al., "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, First Quarter, 2012.
[33] H. Saarimäki et al., "Discrete neural signatures of basic emotions," *Cereb. Cortex*, vol. 26, no. 6, pp. 2563–2573, Apr. 2015.
[34] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affective Comput.*, vol. 7, no. 1, pp. 17–28, First Quarter, 2016.
[35] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans. Cogn. Devel. Syst.*, vol. 14, no. 2, pp. 715–729, Jun. 2022.
[36] J. Li, Z. Zhang, and H. He, "Hierarchical convolutional neural networks for EEG-based emotion recognition," *Cogn. Comput.*, vol. 10, pp. 368–380, 2018.
[37] P. Zhong, D. Wang, and C. Miao, "EEG-Based emotion recognition using regularized graph neural networks," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1290–1301, Third Quarter 2022.
[38] W.-B. Jiang, L.-M. Zhao, P. Guo, and B.-L. Lu, "Discriminating surprise and anger from EEG and eye movements with a graph network," in *Proc. 2021 IEEE Int. Conf. Bioinf. Biomed.*, 2021, pp. 1353–1357.
[39] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affective Comput.*, vol. 11, no. 3, pp. 532–541, Third Quarter, 2020.
[40] Z. Wang, Y. Wang, C. Hu, Z. Yin, and Y. Song, "Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4359–4368, Mar. 2022.
[41] Y. Wang, W.-B. Jiang, R. Li, and B.-L. Lu, "Emotion transformer fusion: Complementary representation properties of EEG and eye movements on recognizing anger and surprise," in *Proc. 2021 IEEE Int. Conf. Bioinf. Biomed.*, 2021, pp. 1575–1578.
[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[43] J. Chen, X. Wang, C. Huang, X. Hu, X. Shen, and D. Zhang, "A large finer-grained affective computing EEG dataset," *Sci. Data*, vol. 10, no. 1, 2023, Art. no. 740.
[44] H. Becker, J. Fleureau, P. Guillotel, F. Wendling, I. Merlet, and L. Albera, "Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources," *IEEE Trans. Affective Comput.*, vol. 11, no. 2, pp. 244–257, Second Quarter, 2020.
[45] X. Hu, F. Wang, and D. Zhang, "Similar brains blend emotion in similar ways: Neural representations of individual difference in emotion profiles," *Neuroimage*, vol. 247, 2022, Art. no. 118819.
[46] N. Koide-Majima, T. Nakai, and S. Nishimoto, "Distinct dimensions of emotion in the human brain and their representation on the cortical surface," *NeuroImage*, vol. 222, 2020, Art. no. 117258. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920307448
[47] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Proc. Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 1–4, 2017.
[48] A. Arjun, A. S. Rajpoot, and M. R. Panicker, "Introducing attention mechanism for EEG signals: Emotion recognition with vision transformers," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, pp. 5723–5726.
[49] A. S. Rajpoot et al., "Subject independent emotion recognition using EEG signals employing attention driven neural networks," *Biomed. Signal Process. Control*, vol. 75, 2022, Art. no. 103547.
[50] L.-M. Zhao, R. Li, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from EEG and eye movement signals: Complementary representation properties," in *Proc. 9th Int. IEEE/EMBS Conf. Neural Eng.*, 2019, pp. 611–614.
[51] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 211–223, Second Quarter, 2012.
[52] B. Sun et al., "Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild," *J. Multimodal User Interfaces*, vol. 10, pp. 125–137, 2016.
[53] K. Guo et al., "A hybrid fuzzy cognitive map/support vector machine approach for EEG-based emotion classification using compressed sensing," *Int. J. Fuzzy Syst.*, vol. 21, pp. 263–273, 2019.
[54] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, "HetEmotionNet: Two-stream heterogeneous graph recurrent neural network for multimodal emotion recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1047–1056.
[55] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Emotion recognition with pre-trained transformers using multimodal signals," in *Proc. 10th Int. Conf. Affect. Comput. Intell. Interaction*, 2022, pp. 1–8.
[56] C. Du et al., "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *Proc. 26th ACM Int. Conf. Multimedia*, New York, NY, USA, 2018, pp. 108–116, doi: 10.1145/3240508.3240528.
[57] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.
[58] D. O. Bos et al., "EEG-based emotion recognition," *The Influence Vis. Auditory Stimuli*, vol. 56, no. 3, pp. 1–17, 2006.
[59] S. B. Eysenck, H. J. Eysenck, and P. Barrett, "A revised version of the psychoticism scale," *Pers. Individual Differences*, vol. 6, no. 1, pp. 21–29, 1985.
[60] T. Schreckenbach et al., "Emotion recognition and extraversion of medical students interact to predict their empathic communication perceived by simulated patients," *BMC Med. Educ.*, vol. 18, no. 1, pp. 1–10, 2018.
[61] H. J. Park and J. H. Lee, "Looking into the personality traits to enhance empathy ability: A review of literature," in *Proc. 22nd Int. Conf. HCI Int. Posters*, Copenhagen, Denmark, Springer, 2020, pp. 173–180.
[62] H. Bao et al., "VLMo: Unified vision-language pre-training with mixture-of-modality-experts," 2021, *arXiv:2111.02358*.
[63] M. Jia et al., "Visual prompt tuning," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Springer, 2022, pp. 709–727.
[64] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.
[65] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.
[66] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[67] Y. Zhang et al., "Free lunch for domain adversarial training: Environment label smoothing," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–38.
[68] A. Gramfort et al., "MEG and EEG data analysis with MNE-Python," *Front. Neurosci.*, vol. 7, no. 267, pp. 1–13, 2013.
[69] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng.*, 2013, pp. 81–84.
[70] L.-C. Shi and B.-L. Lu, "Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, 2010, pp. 6587–6590.
[71] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiol.*, vol. 45, no. 4, pp. 602–607, 2008.
[72] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[73] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv: 1711.05101*.
[74] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process.*, Springer, 2016, pp. 521–529.
[75] X. Cheng et al., "VigilanceNet: Decouple intra-and inter-modality learning for multimodal vigilance estimation in RSVP-based BCI," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 209–217.

[76] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition," *IEEE Trans. Affective Comput.*, vol. 14, no. 3, pp. 2496–2511, Third Quarter 2023.
[77] A. Yoto, T. Katsuura, K. Iwanaga, and Y. Shimomura, "Effects of object color stimuli on human brain activities in perception and attention referred to EEG alpha band response," *J. Physiol. Anthropol.*, vol. 26, no. 3, pp. 373–379, 2007.
[78] W. Klimesch, M. Doppelmayr, H. Russegger, T. Pachinger, and J. Schwaiger, "Induced alpha band power changes in the human EEG and attention," *Neurosci. Lett.*, vol. 244, no. 2, pp. 73–76, 1998.
[79] R. L. van den Brink, P. R. Murphy, and S. Nieuwenhuis, "Pupil diameter tracks lapses of attention," *PLoS One*, vol. 11, no. 10, 2016, Art. no. e0165274.
[80] S. Haro, H. M. Rao, T. F. Quatieri, and C. J. Smalt, "EEG alpha and pupil diameter reflect endogenous auditory attention switching and listening effort," *Eur. J. Neurosci.*, vol. 55, no. 5, pp. 1262–1277, 2022.

Wei-Bang Jiang received the bachelor's degree in computer science and technology from Zhiyuan College, Shanghai Jiao Tong University, Shanghai, China, in 2021. He is currently working toward the PhD degree in computer science and technology with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His research focuses on affective computing, brain-computer interface, and machine learning.

Xuan-Hao Liu received the bachelor's degree in theoretical and applied mechanics from the School of Aeronautics and Astronautics, Sun Yat-sen University, Guangzhou, China, in 2022. He is currently working toward the PhD degree in computer science and technology with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His research focuses on affective computing, brain-computer interface, and machine learning.

Wei-Long Zheng (Member, IEEE) received the bachelor's degree in information engineering from the Department of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, in 2012, and the PhD degree in computer science from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2018. In 2018–2020, he was a research fellow with the Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts and a postdoc associate with the Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Boston, Massachusetts, in 2020–2021. He is currently an associate professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He received the ACM Multimedia 2022 Top Paper Award, the 2021 Best of IEEE Transactions on Affective Computing Paper Collection, and the IEEE Transactions on Autonomous Mental Development Outstanding Paper Award from IEEE Computational Intelligence Society, in 2018. He is currently the associate editor of *IEEE Transactions on Affective Computing*. His research focuses on computational neuroscience, affective computing, brain-computer interface, machine learning, and pattern recognition.

Bao-Liang Lu (Fellow, IEEE) received the BS degree in instrument and control engineering from the Qingdao University of Science and Technology, Qingdao, China, in 1982, the MS degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 1989, and the Dr Eng degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1994. He was with the Qingdao University of Science and Technology from 1982 to 1986. From 1994 to 1999, he was a Frontier researcher with the Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research (RIKEN), Nagoya, Japan, and a research scientist with the RIKEN Brain Science Institute, Wako, Japan, from 1999 to 2002. Since 2002, he has been a full professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He received the IEEE Transactions on Autonomous Mental Development Outstanding Paper Award from IEEE Computational Intelligence Society, in 2018, the Best of IEEE Transactions on Affective Computing Paper Collection, in 2021, the ACM Multimedia 2022 Top Paper Award, and 2022 APNNS (Asia Pacific Neural Network Society) Outstanding Achievement Award. He was the president of the Asia Pacific Neural Network Assembly and the general chair of the 18th International Conference on Neural Information Processing, in 2011. He is currently the associate editor of *IEEE Transactions on Affective Computing* and *Journal of Neural Engineering*. His current research interests include brain-like computing, deep learning, affective computing, and brain-computer interface.