CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/154335/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Du, Xiaobing, Deng, Xiaoming, Qin, Hangyu, Shu, Yezhi, Liu, Fang, Zhao, Guozhen, Lai, Yu-Kun , Ma, Cuixia, Liu, Yong-Jin and Wang, Hongan 2023. MMPosE: Movie-induced multi-label positive emotion classification through EEG signals. IEEE Transactions on Affective Computing 14 (4) , pp. 2925-2938. 10.1109/TAFFC.2022.3221554

Publishers page: http://dx.doi.org/10.1109/TAFFC.2022.3221554

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

MMPosE: Movie-induced Multi-label Positive Emotion Classification through EEG Signals

Xiaobing Du, Xiaoming Deng, Hangyu Qin, Yezhi Shu, Fang Liu, Guozhen Zhao, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, *Senior Member, IEEE,* and Hongan Wang, *Senior Member, IEEE*

Abstract—Emotional information plays an important role in various multimedia applications. Movies, as a widely available form of multimedia content, can induce multiple positive emotions and stimulate people's pursuit of a better life. Different from negative emotions, positive emotions are highly correlated and difficult to distinguish in the emotional space. Since different positive emotions are often induced simultaneously by movies, traditional single-target or multi-class methods are not suitable for the classification of movie-induced positive emotions. In this paper, we propose *TransEEG*, a model for multi-label positive emotion classification from a viewer's brain activities when watching emotional movies. The key features of *TransEEG* include (1) explicitly modeling the spatial correlation and temporal dependencies of multi-channel EEG signals using the Transformer structure based model, which effectively addresses long-distance dependencies, (2) exploiting the label-label correlations to guide the discriminative EEG representation learning, for that we design an Inter-Emotion Mask for guiding the Multi-Head Attention to learn the inter-emotion correlations, and (3) constructing an attention score vector from the representation-label correlation matrix to refine emotion-relevant EEG features. To evaluate the ability of our model for multi-label positive emotion classification, we demonstrate our model on a state-of-the-art positive emotion database CPED. Extensive experimental results show that our proposed method achieves superior performance over the competitive approaches.

Index Terms—Multi-channel EEG; positive emotions; affective computing; multi-label classification; Transformer encoder

1 Introduction

Movies enable people to experience emotions through hearing and vision, thereby generating sensory pleasure. "It is only in the mysterious equations of love that any logic or reasons can be found." When John Nash confided in his wife at the Nobel Prize ceremony, we are all deeply touched by the pure love, steadfast companionship and respect for science in the story told by the movie *A Beautiful Mind*. As this example shows, movies are created with the intended purpose to evoke an emotional response of the viewer. As stated in [1], "Movies dazzle us, entertain us, educate us, and delight us".

- *X.B. Du, X.M. Deng, H.Y. Qin are with the Beijing Key Laboratory of Human Computer Interactions, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China, and University of Chinese Academy of Sciences, Beijing, China. E-mail: {duxiaobing16, qinhangyu20}@mails.ucas.ac.cn, xiaoming@iscas.ac.cn.*
- *Y.Z. Shu, F. Liu and Y.-J. Liu are with BNRist, MOE-Key Laboratory of Pervasive Computing, the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: {shuyz19@mails., lfang@, liuyongjin@}tsinghua.edu.cn.*
- *Y.-K. Lai is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, Wales, UK. E-mail: laiy4@cardiff.ac.uk*
- *G.Z. Zhao is with the CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China, and also with the Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: zhaogz@psych.ac.cn.*
- *C.X. Ma and H.A. Wang are with the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China, University of Chinese Academy of Sciences, Beijing, China, and also with the Beijing Key Laboratory of Human Computer Interactions, International Joint Laboratory of Artificial Intelligence and Emotional Interaction, Beijing 100190, China. E-mail: {cuixia, hongan}@iscas.ac.cn*
- *Y.-J. Liu, C.X. Ma and G.Z. Zhao are the corresponding authors.*

Movies may induce emotion through cognitive causality, for instance, cognitively appreciating injustice tends to give rise to anger, while loss tends to lead to sadness [2]. According to the concept of mood[1] management proposed by [3], almost all movie selections are designed for the outcome of pleasure or increased positive effects. Even horror movies can be chosen for the same purpose, like enjoying the pleasure of releasing or resolving from the tension. Movies can bring short-term effects to the viewer, and promote people's long-term changes in behaviors and characteristics [4]. According to the B. L. Fredrickson's broaden-and-build theory of positive emotion [5], one's positive emotion can be multiplied into an upward spiral, and human beings are able to transform current positive emotions into positive resources for future needs during hard times. So the positive emotions have certain positive effects on human life, which may not only help identify our role models and mentors, but also help raise the character strengths. The positive emotions induced by movie can stimulate people's pursuit of a better life, including but not limited to building their own characters, emphasizing self-acceptance, and improving current life satisfaction [6]. Therefore, it has great relevance to explore the positive emotions and emotional cognition perceived by the viewer while watching movies.

In our study, we pay attention to positive emotions due to their unique cognitive functions. Different from negative emotions, positive emotions are highly correlated and indistinguishable in the emotional space [7], [8]. Moreover, plots in the movie have a temporal continuity, so that the

1. "Emotion" and "mood" are often used interchangeably. The "mood" here is the same as "emotion".

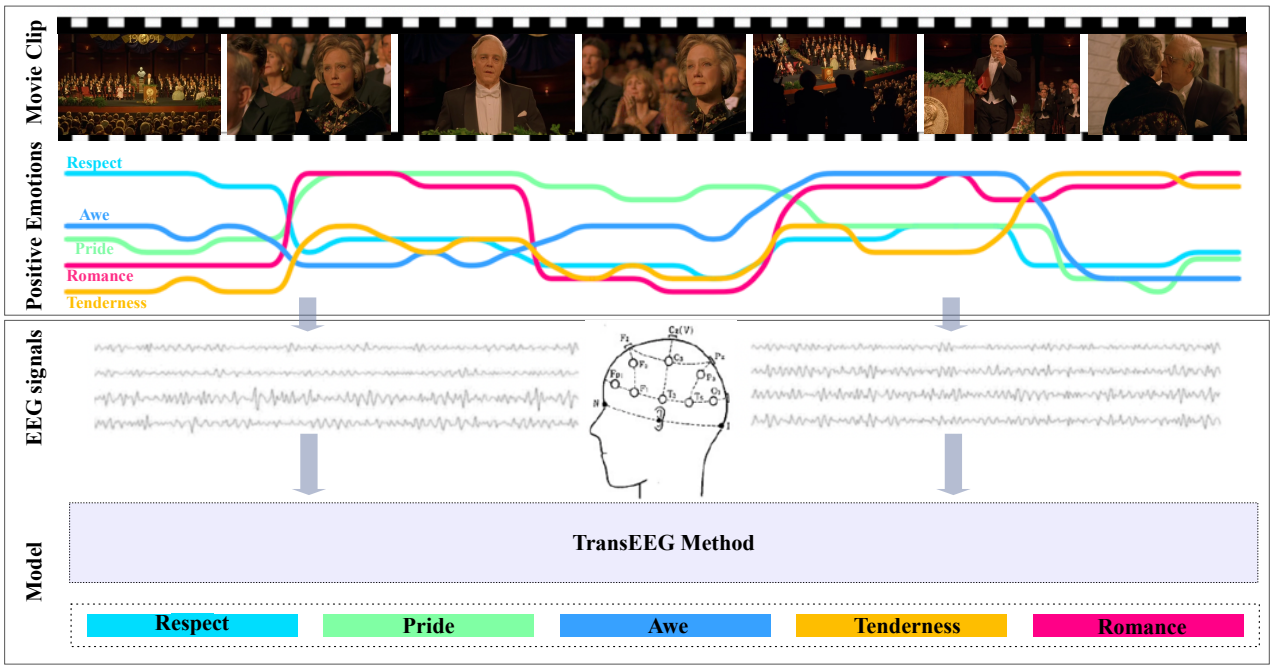

Fig. 1. Given EEG signals captured during movie watching as input, our *TransEEG* model predicts the evoked multiple positive emotions for each movie clip. Here, we use a clip from the movie *A Beautiful Mind* to show the procedure of analyzing movie-induced positive emotions. Five positive emotions were evoked by this movie clip, and their intensities changed dynamically over time as shown by the plotted curves.

multiple co-occurring emotions with potential correlations are always induced in viewers. Therefore, the traditional single-label emotion classification model (i.e., assigning one label to a movie clip) is oversimplified and not suitable for fine-grained discrete positive emotion analysis. In this paper, we recognize the movie-induced positive emotions by formulating the problem as a multi-label classification task. Furthermore, multi-label classification, which aims to identify coexisting subsets of emotions present in movie clips (e.g., entertainment, respect and pride), has gained widespread attention due to its wide range of potential applications, such as personalized movie recommendation, movie retrieval, and human robot interaction, etc. In this paper, we propose a novel *TransEEG* model which is built based on the Transformer structure to accomplish multi-label positive emotion classification.

Traditionally, positive emotions are associated with one expression, i.e. the Duchenne smile [9], which is characterized by the increased activity in zygomatic major and orbicularis oculi muscles. However, recent studies have shown that positive emotions are not necessarily associated with smiles. And Campos et al. [7] examined the expressive display patterns of eight positive emotions and found that amusement, joy, contentment, love, and pride resulted in smiles that varied in intensity, whereas awe and interest did not result in smiles. Therefore, it is challenging for positive emotion recognition based on multimedia content [10]. Although existing facial expression recognition methods [11], [12] in the field of computer vision can achieve outstanding performance, significant difficulties still exist for recognizing fine-grained positive emotions through the facial expressions of viewers or characters in a movie.

In recent years, many studies have been attempted to address whether discrete positive emotions can be differentiated via related emotional responses [13]. Moreover, several works have shown that positive emotions can be associated with divergent patterns of physiological activities [14], and electroencephalography (EEG) signals have been shown to be effective for recognizing discrete positive emotions [8], [15], [16], [17], [18]. Therefore, we recognize positive emotions induced by movie via viewer's brain activities (i.e., EEG signals) in this paper, which is validated on the state-of-the-art Chinese Positive Emotion Database (CPED) [18]. This database consists of effective positive emotion-inducing materials (i.e., movie clips) and biological signals (i.e., EEG signals) collected by eliciting emotions during watching movie clips. Each EEG trial in the database is annotated with a binary label vector of length 9, corresponding to the state of 9 positive emotions (i.e., Friendship, Romance, Tenderness, Respect, Pride, Awe, Gratitude, Amusement, and Craving). Specifically, $0/1$ in the binary label vector means that the emotion is absent/present in the movie clip. Moreover, these 9 discrete positive emotions have been proved to be the most commonly experienced feelings in daily life, and could be identified effectively via EEG signals [8], [16], [17], [18].

Recognizing positive emotions from the perspective of predicting the viewer's physiological response to a movie can be a testbed for movie emotion analysis. In practice, many research areas, including human-computer interaction, automated labeling systems and personalized movie recommendation systems, would benefit a lot if modern artificial intelligence systems had the ability to effectively understand the positive emotions conveyed by movie contents. Recent studies on EEG-based emotion recognition have been developed, and the deep neural networks are superior in extracting EEG emotion-relevant features. Moreover, the representative neural networks include Convolutional Neural Networks (CNN) [19], Graph Convolutional

Neural Network (GCN) [20], Long Short-Term Memory (LSTM) [21], and attention mechanism [22] have demonstrated that learning the spatial correlation and temporal dependencies of channels is essential for extracting discriminative EEG features. Especially, Transformer [23] is a self-attention based architecture with the ability to establish long-distance dependencies, which emerged as the preferred model in sequential data processing. In this paper, to effectively learn the long-distance sequential correlation of EEG channels and temporal dependencies, we design our model based on the multi-layer Transformer structure. The main contributions of our work can be summarized as follows:

- We propose a Transformer-based model *TransEEG* to address the multi-label positive emotion classification task. The spatial-temporal encoding module is designed based on the dual-stream parallel Transformer encoders to extract the spatial-temporal EEG features. Specifically, the spatial encoder learns the inter-channel correlations and the temporal encoder captures the temporal dependencies of a time sequence of EEG channels. Afterwards, the spatial-temporal EEG features can be obtained by integrating the outputs of the two Transformer encoders.
- Exploiting the label-label correlations to assist in extracting discriminative emotion-relevant features. An Inter-Emotion Mask is designed based on the co-occurrence of emotions in the training set to guide label-label correlation modeling in Transformer. After that, the representation-label correlation matrix is generated and utilized to generate the attention scores focusing on EEG features. And the attention mechanism attempts to refine the emotion-relevant EEG representations.
- The *TransEEG* learns the discriminative EEG representation by capturing spatial-temporal information and leveraging the label-label correlations. Extensive experiments indicate that our method achieves competitive performance against other classical multi-label classification approaches and the state-of-the-art deep neural networks in the field of EEG based emotion recognition.

TransEEG is proposed to predict the multi-label positive emotions induced by movies from multi-channel EEG signals. Our work shows that it is a successful attempt to effectively recognize positive emotions induced by movies through viewers' physiological responses. An example is illustrated in Fig. 1.

2 Related Work

Our work is related to several research areas, including emotion models and discrete positive emotions, multi-label emotion classification, EEG-based emotion recognition and long sequence modeling.

2.1 Emotion Models and Discrete Positive Emotions

Emotion models have been widely studied in psychology, and there are generally two basic models [8], [17], [22], [24]: dimensional models and discrete models. Dimensional models describe emotion states in a 2D or 3D continuous space, such as the classic valence-arousal (VA) model or the valence-arousal-dominance (VAD) model. Discrete models on the other hand describe emotion states using a limited number of basic emotions. If dimensional models are used to characterize positive emotions, it is difficult to distinguish them from each other due to high correlations [25], i.e., their coordinates in continuous space are very close and clustered in a small region. Therefore, discrete models are preferred to describe positive emotions.

Many studies have been proposed to use discrete models to categorize different positive emotions, mostly based on subjective reports [26], [27]. For example, Watson et al. [28] introduced ten positive emotions using the well-established Positive Affect and Negative Affect Scale (PANAS). Fredrickson [26] presented another set of ten positive emotions, which were suggested to be representative of the emotions in daily life. Zhang et al. [18] employed movie clips to elicit sixteen discrete positive emotions and these emotions were further categorized into four main emotion categories (Empathy, Fun, Creativity and Esteem) in CPED [18]. In this paper, we adopt the state-of-the-art CPED database, since it is the first standardized video-based positive emotion database. After emotion clustering, we select nine most representative positive emotion categories. Specifically, Tenderness, Gratitude and Romance are selected from Empathy, Amusement and Friendship from Fun, Awe and Craving from Creativity, and Respect and Pride from Esteem.

2.2 Multi-label Emotion Classification

The multi-label classification (e.g., the positive emotion subset can be predicted to occur in one movie clip) is desired in many areas of research [29], especially in the field of affective computing. For example, in the area of video emotion recognition, Zhang et al. [30] proposed a multi-modal seq2set (MMS2S) approach to address the challenge in multi-label emotion detection in video clips. Kostiuk et al. [31] extended the CAL500 database by including music videos, and tried to address the classification problem of multi-label emotions in music videos. For text sentiment analysis, Fei et al. [32] presented a latent emotion memory (LEM) network to learn the latent emotion distribution and recognize multi-label emotions in a sentence. Aiming at facial expression recognition, Li and Deng [33] proposed a new deep manifold learning network to learn discriminative features of multi-label expressions. Moreover, Zhang et al. [34] focused on multi-modal emotion recognition in a multi-label scenario.

In recent years, works on positive emotion analysis have received much attention, since the related studies have shown that positive emotions have significant effects on human life [6], [26] and can help people cope with negative events [35]. Furthermore, recognizing positive emotions via viewer's EEG signals has also been investigated [8], [16], [17], [18]. For instance, Hu et al. [16] used EEG spectral powers to classify the discrete positive emotions. Moreover, they further reported recognizable discrete positive emotions using the functional Near-Infrared Spectroscopy (fNIRS) signal [8]. Zhao et al. [17] adopted linear and non-linear models to recognize four positive emotion categories

using EEG power features. However, the above studies are based on single-label classification methods, which performs coarse-grained positive emotion classification with fewer positive emotion categories. Furthermore, to the best of our knowledge, it is barely investigated to design a deep learning model based on a multi-label classification task to analyze fine-grained positive emotions by EEG signals.

Moreover, compared to movie emotion recognition based on multimedia content [36], [37], the advancement of work on movie emotion analysis via viewer's physiological signals [38], [39], [40], [41] provides a foundation for studies on movie induced positive emotion recognition. However, few relevant studies on movie-induced positive emotions have been conducted in the multimedia emotion analysis community, as the accurate classification of discrete positive emotions is a significant challenge for single-label classification methods. Therefore, in this paper, we employ a multi-label classification method to address the challenge of movie-induced positive emotions analysis.

2.3 EEG-based Emotion Recognition

EEG can faithfully reflect different emotions by directly capturing brain activities via the electrodes attached to the scalp. Moreover, processing EEG signals with high temporal resolutions is a reliable way to identify real emotions. Many methods have been proposed to deal with EEG-based emotion recognition [21], [22], [24], in which deep learning models achieved the superior results. Several representative deep models are as follows. Zheng and Lu [42] have attempted to apply the Deep Belief Network (DBN) to extract high-level emotional features, and the results demonstrate that deep models can extract highly effective EEG features. Ma et al. [43] proposed a multimodal residual LSTM (MM-ResLSTM) network to capture the temporal information for enhancing multimodal emotion recognition task with EEG and other physiological signals. Du et al. [21] designed an efficient deep neural network to enhance the emotion recognition performance by learning the correlation between EEG channels. Jia et al. [22] proposed a novel spatial-spectral-temporal based attention 3D dense network to use different EEG features and the discriminative local patterns among features for emotion recognition. Learning the spatial correlation between channels and extracting EEG spatial features has received significant attention at present. Specifically, Song et al. [20] proposed a novel dynamic graph convolutional neural network (DGCNN) to learn the functional relation between each pair of two channels, and Zhang et al. [44] designed a sparse DGCNN model to improve the DGCNN by applying a sparseness constraint on the weight graph.

Although above works have demonstrated the importance to learn the relationship between EEG channels and temporal dependencies for emotion recognition, few works can encode EEG signals from spatial and temporal dimensions simultaneously. In this paper, we strengthen the EEG emotion recognition method by simultaneously encoding EEG signals in spatial-temporal space, and construct our model based on the Transformer encoder structure that has a prominent ability to model long sequence data.

2.4 Long Sequence Modeling

Transformer [23] is a self-attention based network structure, which has the outstanding ability to address long-distance dependencies in natural language processing and computer vision, such as the BERT [45] and ViT [46]. Due to the superiority of the multi-head attention mechanism for long-distance dependency learning, Transformer-based approaches have also been proposed for processing EEG signals recently. Wang et al. [47] proposed a Transformer-based model to hierarchically extract the discriminative spatial features from the electrode level to the brain region level. Guo et al. [48] combined depthwise convolution and Transformer encoders to explore the dependencies of emotion recognition on each EEG channel. Sun et al. [49] constructed multiple Transformer-based models for motor imaginary EEG classification by learning the correlation of time series signals.

The studies mentioned above have shown that Transformer is a powerful model to extract discriminative EEG features. Nevertheless, they ignored combining spatial and temporal encoding of EEG data to learn more discriminative features. In this paper, to better model the long-distance sequential correlation of EEG channels and temporal dependencies, we construct our model based on the multi-layer Transformer structure.

3 The Proposed TransEEG Model

3.1 Overview

Our deep neural network architecture for movie-induced multi-label positive emotion classification, called *TransEEG*, mainly consists of three key modules (see Fig. 2): *Spatial-temporal encoding module*, *Label correlation learning module*, and *Correlation-guided representation learning module*. The spatial-temporal encoding module and the label correlation learning module are both built upon the multi-layer Transformer encoder structure. The spatial encoder utilizes the effective multi-head attention in Transformer to learn the inter-channel correlations of the input EEG channel sequence. We also adopt the temporal encoder to learn the key temporal features of time-sequential EEG data (Section 3.3.2). Moreover, we design an Inter-Emotion Mask based on the emotion label co-occurrence in the training set to learn label-label correlation (Section 3.3.3). Then, we use the correlation guided EEG representation learning module to extract emotion-relevant features by leveraging the spatial-temporal EEG features as well as the learned representation-label correlations (Section 3.3.4).

3.1.1 The Motivation of Inter-Emotion Mask

Related studies have shown that there are certain correlations existing among positive emotions [16], [17], [18]. Correspondingly, positive emotions in the CPED database are also correlated with each other. An example is shown in Fig. 3, in which the Pearson correlations were calculated in the 9 positive emotion categories corresponding to the movie clip of *An interview with Qian Xuesen*. Therefore, it is crucial to investigate how to leverage the label-label correlations for the multi-label positive emotion classification procedure. In this paper, we capture label-label correlations

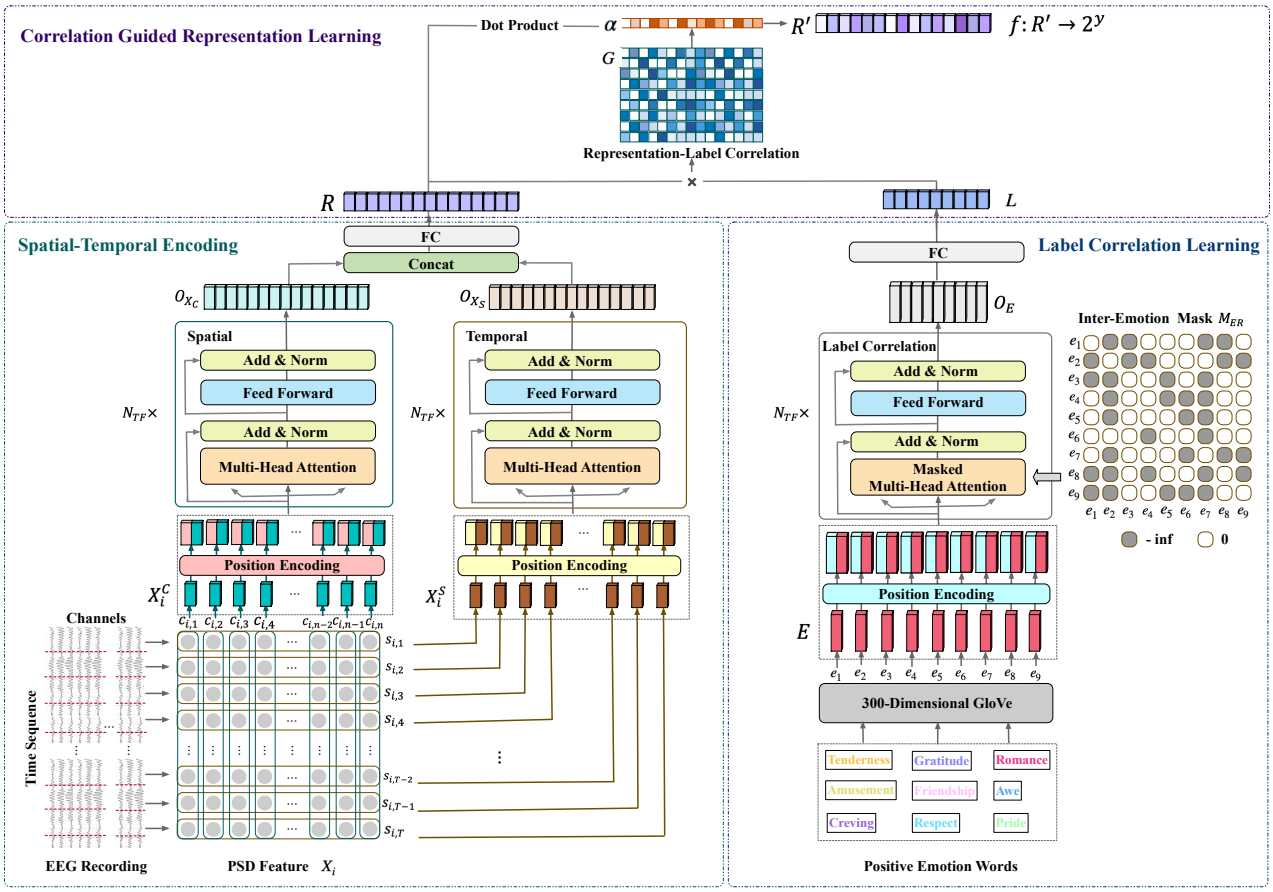

Fig. 2. The framework of our proposed *TransEEG* model, which is based on the Transformer encoder structure. Our model consists of three main modules: the Spatial-temporal encoding module, the Label correlation learning module, and the Correlation guided representation learning module. The spatial-temporal encoding module extracts the spatial features $\mathbf{O}_{X_C}$ and the temporal features $\mathbf{O}_{X_S}$, and the spatial-temporal EEG features $\mathbf{R}$ can be obtained by concatenating spatial and temporal features. The label correlation learning module learns the label correlation representation $\mathbf{L}$ under the guidance of the Inter-Emotion Mask $\mathbf{M}_{ER}$. To focus on emotion-related features with the attention mechanism, the representation-label correlation matrix $\mathbf{G}$ was first generated by simply multiplying $\mathbf{R}$ by $\mathbf{L}$. Then the attention score vector α is calculated from $\mathbf{G}$. Subsequently, the correlation guided representation $\mathbf{R}^{'}$ is generated from the dot product of $\mathbf{R}$ and α. Finally, $\mathbf{R}^{'}$ is used to predict positive emotion labels.

through multi-head attention in the Transformer to guide the EEG representation learning.

Furthermore, masks in the Transformer are applied to mask unattended elements in self-attention [45], [50], [51], which are flexible and convenient to be designed. In this paper, we design an Inter-Emotion Mask based on the co-occurrence of emotions in the training dataset to enhance label-label correlation learning in multi-head attention. Since the attention mechanism can assist in focusing more on the relevant features, an attention score vector is also learned from the representation-label correlation matrix to refine the emotion-relevant EEG features. Specifically, for the aims of obtaining attention score vector, we integrate spatial-temporal representation and label correlations to construct the representation-label correlation matrix. Eventually, an attention score vector is achieved, and the correlation guided representation is successively calculated for positive emotion classification.

3.1.2 Problem Formulation

Given a training set $\mathcal{X} = \{(\mathbf{X}_i, Y_i)\}_i^n$ of multi-label positive emotion classification data, where $\mathbf{X}_i \in \mathcal{X}$ is the i-th instance consisting of the power spectral density (PSD) of the sequential EEG signals [18] and $Y_i \subseteq \mathcal{Y}$ is its corresponding labels, and $\mathcal{Y}$ represents the label space. In this paper, the multi-label positive emotion classification task aims to learn a predictive function f to predict a subset of the possible labels for an instance. Specifically, $\mathbf{X}_i \in \mathbb{R}^{T \times N \times D}$ where T is the length of the time series of an instance, and we set $T = 30$ in our experiments. N is the number of EEG channels, which is 30 according to the used EEG recording device. $D = 5$ is the dimension of the extracted PSD features on a single channel through a sliding window.

Specifically, a sequence of EEG channels of length N is composed of the PSD features of all channels $\mathbf{X}_i^C = \{c_{i,1}, c_{i,2}, ..., c_{i,n}, \ldots, c_{i,N}\}$, which is the input for the spatial encoding module, and $c_{i,N} \in \mathbb{R}^{d_c}$ represents the PSD features of the N-th channel in i-th instance, and $d_c = 150$ is the feature dimension. Moreover, an input EEG time sequence of length T can be represented as $\mathbf{X}_i^S = \{s_{i,1}, s_{i,2}, ..., s_{i,t}, \ldots, s_{i,T}\}$, where $s_{i,T} \in \mathbb{R}^{d_s}$ denotes the PSD features of the T-th time step in i-th instance, and $d_s = 150$ is the feature size. Denote $Y_i = \{y_i^1, y_i^2, ..., y_i^k, ..., y_i^l\}$ to be the label space with l labels (l is the number of positive emotion categories, and set to 9 in this paper), where $y_i^k \in \{0, 1\}$ means that the k-th emotion e_k is irrelevant (i.e. $y_i^k = 0$) or relevant (i.e. $y_i^k = 1$)

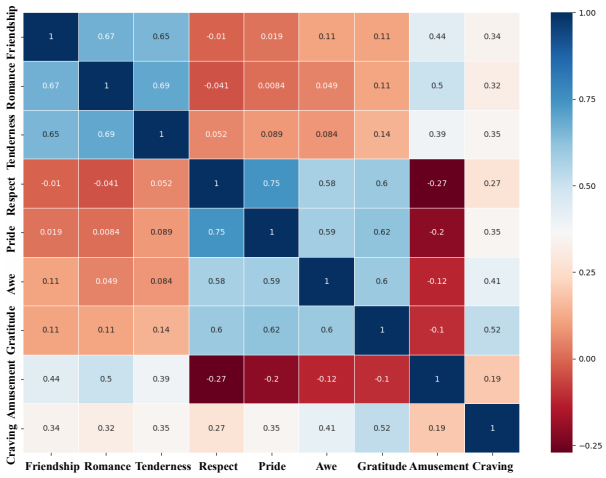

Fig. 3. An illustration of the correlations among 9 positive emotions that occurred in the movie clip from *An interview with Qian Xuesen* in the CPED database. Correlations are visualized in color: blue represents higher correlation, and red represents lower correlation or negative correlation.

to the instance $\mathbf{X}_i$. Compared to single-label classification where only one label is associated with $\mathbf{X}_i$, the multi-label positive emotion prediction function can be formulated as $f : \mathcal{X} \rightarrow 2^{\mathcal{Y}}$, which assigns a subset of possible class labels Y to an instance $\mathbf{X}_i$, where $1 \leqslant |Y| \leqslant l$. The Inter-Emotion Mask $\mathbf{M}_{ER}$ can be represented as $\mathbf{M}_{ER} \in \mathbb{R}^{l \times l}$, and a sequence of emotion label embedding of length l can be represented as $\mathbf{E} = \{e_1, e_2, ..., e_l\} \in \mathbb{R}^{l \times d_l}$ where $d_l = 300$ is the dimension of the emotion word embedding encoded by GloVe [52].

After we obtain the EEG spatial-temporal features $\mathbf{R}$ and the label correlation representation $\mathbf{L}$, the representation-label correlation matrix $\mathbf{G}$ can be calculated by multiplying $\mathbf{R}$ by $\mathbf{L}$. Afterwards, the attention score vector α calculated from $\mathbf{G}$ is used to multiply by the spatial-temporal features $\mathbf{R}$ to gain the correlation guided representation $\mathbf{R}^{'}$. Thereby multi-label positive emotion prediction function can be updated to $f : \mathbf{R}^{'} \rightarrow 2^{\mathcal{Y}}$, which aims to enforce the model to accurately predict the relevant labels for each training instance by minimizing the loss function.

3.2 Time-Frequency EEG Features

EEG PSD is one of the most widely used EEG features in EEG-based emotion recognition [18], [42], [53]. In CPED database, the PSD features extracted from multi-channel EEG signals are provided for our spatial-temporal encoding module as input handcraft EEG features. Specifically, the short-time Fourier transform (STFT) is used to extract classical PSD features using a 2-seconds sliding Hamming window with a 50% overlap across the five sub-frequency bands, i.e., δ band (1-3Hz), θ band (4-7Hz), α band (8-13Hz), β band (14-30Hz), and γ band (31-50Hz). There are a total of 30 channels[2] of the recorded EEG signals, while the PSD features are captured every 1 time step on 5 frequency bands and the dimension of the feature input to the temporal encoder is also 150 ($150 = D \times N$). Moreover, the length of the time sequence in an instance is 30, thus the size of PSD feature input to the spatial encoder is 150 ($150 = D \times T$).

In our work, we organize PSD features according to channel sequence and time sequence simultaneously to facilitate the extraction of spatial-temporal EEG features via the spatial-temporal encoding module.

2. The 30 electrodes are selected from a 32-electrode Neuroscan Quik-Cap (https://compumedicsneuroscan.com/products/caps/quik-cap/) according to the international 10–20 system. The remaining two electrodes are used as reference electrodes.

3.3 Model Architecture

Since each layer of the multi-layer Transformer encoder used in this work has the same network structure, we briefly describe the general procedure of data processing within the first layer of a Transformer encoder. In this paper, we utilize the Transformer encoder to capture the sequential dependencies from a sequence of EEG data $\mathbf{X}_i$.

3.3.1 *Preliminary Knowledge*

To enable the Transformer encoder to exploit the sequential relationship of the EEG data, we append the absolute position information of the sequential data to $\mathbf{X}_i$. And consequently the input (i.e. $\mathbf{X}_i^p$) of the Transformer encoder is obtained, which can be formulated as:

$$\mathbf{X}_i^p = \mathbf{X}_i + PE_{pos} \tag{1}$$

where PE_{pos} denotes the positional encodings. Although many positional encoding functions are available, we choose the popularly used sine and cosine functions of different frequencies to encode the position information [54]. Thereafter, we omit the subscript i in $\mathbf{X}_i^p$ for the sake of simplicity.

Then we elaborate on the two key sub-layers of Transformer encoder, namely a multi-head attention layer (MHA) and a simple position-directed fully connected feed-forward network (FFN). The multi-head attention is achieved by k self-attentions, which able to attend to information from different representation sub-spaces at different positions. Please refer to [54] for more details. In the process, an input sequence $\mathbf{X}^p \in \mathbb{R}^{N \times d}$ can be projected to query matrix $\mathbf{Q} \in \mathbb{R}^{N \times d_k}$, key matrix $\mathbf{K} \in \mathbb{R}^{N \times d_k}$ and value matrix $\mathbf{V} \in \mathbb{R}^{N \times d_v}$. Note that $d_k = d_v = d$ in this paper. The outputs of all k heads are then concatenated and projected to $\mathbf{O}_a$, which is the output of multi-head attention layer with the same size as X^p. The formulation can be briefly represented as:

$$\mathrm{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, ..., \mathrm{head_k})\mathbf{W}^o \tag{2}$$

$$\mathrm{head_i} = \mathrm{Attention}(\mathbf{Q}\mathbf{W}_i^q, \mathbf{K}\mathbf{W}_i^k, \mathbf{V}\mathbf{W}_i^v) \tag{3}$$

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}(\frac{(\mathbf{Q}\mathbf{K}^T)}{\sqrt{d_k}})\mathbf{V} \tag{4}$$

where $\mathbf{W}_i^q \in \mathbb{R}^{d_k \times d_k^{'}}$, $\mathbf{W}_i^k \in \mathbb{R}^{d_k \times d_k^{'}}$, $\mathbf{W}_i^v \in \mathbb{R}^{d_v \times d_v^{'}}$ and $\mathbf{W}^o \in \mathbb{R}^{kd_v^{'} \times d_v}$ are the linear project matrices. Here we use 6 parallel attention heads, and set $d_k^{'} = d_v{'} = d/k$.

After the multi-head attention layer, a feed-forward network layer is applied to each position separately. The feed-forward network layer consists of two linear transformations with a ReLU activation in between. Thus the procedure of feed-forward network can be formulated as:

$$\mathrm{FFN}(\mathbf{O}_a^{'}) = \max(0, \mathbf{O}_a^{'}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{5}$$

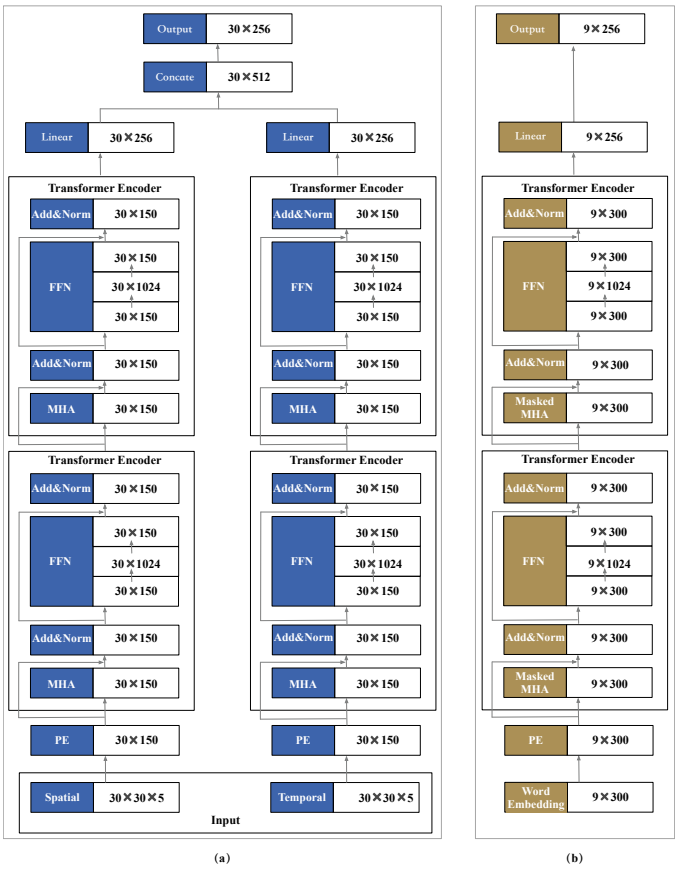

Fig. 4. The structure diagram of spatial-temporal encoding module and label correlation learning module.The left of each block in this diagram is the type of layers and the right is the dimension of its output tensor.

where $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$ and $\mathbf{b}_2$ are the learnable parameters. There is a LayerNorm after each multi-head attention layer and feed-forward network layer, and both LayerNorms are residual connected. The output of LayerNorm (i.e. $\mathbf{O}_a^{'}$) can be obtained by:

$$\mathbf{O}_a^{'} = \mathrm{LayerNorm}(\mathbf{O}_a + \mathbf{X}^p) \tag{6}$$

The output $\mathbf{O}^1$ of the first layer of Transformer encoder can be calculated by:

$$\mathbf{O}^1 = \mathrm{LayerNorm}(\mathrm{FFN}(\mathbf{O}_a^{'}) + \mathbf{O}_a^{'}) \tag{7}$$

Afterwards, $\mathbf{O}^1$ acts as the input of the second layer of the Transformer encoder. By analogy, we can obtain the final output $\mathbf{O}_X \in \mathbb{R}^{N\times d}$ of multi-layer Transformer encoder.

3.3.2 Spatial-Temporal Encoding Module

We leverage both spatial and temporal information to learn an effective spatial-temporal EEG representation. We design a spatial-temporal encoding module to extract spatial-temporal features, and the features are provided as input to discriminative emotion-relevant EEG feature learning. Moreover, the spatial-temporal encoding module consists of two sub-modules: the spatial encoder and the temporal encoder, and its structure is shown in Fig. 4(a).

Spatial Encoder utilizes the Transformer encoder to capture spatial dependencies from the sequence of EEG channels. The input data of the spatial encoder is $\mathbf{X}_i^C \in \mathbb{R}^{N\times d_c}$. After the positional encoding process, multi-head attention in the spatial encoder models the long-distance interactions between channels. Following the multi-layer Transformer processing, we can obtain the output feature $\mathbf{O}_{X_C} \in \mathbb{R}^{N\times d_c}$ of spatial encoder, which acts as the EEG spatial encoding.

Temporal Encoder aims to learn the temporal dependencies of EEG time series data, and $\mathbf{X}_i^S \in \mathbb{R}^{T\times d_s}$ denotes the inputs of a time sequence. After the multi-layer Transformer processing through multi-head attention and feed-forward network, the output $\mathbf{O}_{X_S} \in \mathbb{R}^{T\times d_s}$ of temporal encoder is the ultimate EEG temporal encoding.

Spatial-Temporal Encoding can be obtained by concatenating $\mathbf{O}_{X_C}$ and $\mathbf{O}_{X_S}$.

$$\mathbf{R} = linear(\mathrm{Concat}(\mathbf{O}_{X_C}, \mathbf{O}_{X_S})) \tag{8}$$

where $\mathbf{R} \in \mathbb{R}^{N\times d_{st}}$, which denotes the spatial-temporal encoding of EEG signals, and $d_{st} = d_s = d_c$. After that, $\mathbf{R}$ is used for learning the correlation guided representation based on the representation-label correlation matrix in the next procedure.

3.3.3 Label Correlation Learning Module

Since emotions with stronger correlation tend to have more similarities in their emotional experience, we adopt the co-occurrence of positive emotions in the movie clips as inter-emotion correlations. In this paper, we propose a label correlation learning module based on Transformer encoder structure to further learn the correlations between emotions. The architecture details of this module are shown in Fig. 4 (b). Furthermore, to facilitate multi-head attention in Transformer encoder to learn the dependencies between labels, we design a novel Inter-Emotion Mask according to the emotion co-occurrence matrix.

Inter-Emotion Mask is designed for the masked multi-head attention in Transformer (see Fig. 2), and it aims to mask the unattended emotions in self-attention. The Inter-Emotion Mask is constructed according to the inter-emotion correlation matrix $\mathbf{P}$, and the dependencies of emotion labels can be represented by $p(e_i|e_j)$ (i.e., the conditional probability of the occurrence of the emotion e_i with e_j.). To construct $\mathbf{P}$, we firstly calculate the number of occurrences of emotion pairs to obtain the co-occurrence matrix $\mathbf{M} \in \mathbb{R}^{l\times l}$, where $M_{i,j}$ denotes the number of co-occurrences of e_i and e_j. Then, we calculate the correlation matrix $\mathbf{P} = \{P_{i,j}\} \in \mathbb{R}^{l\times l}$ based on the emotion co-occurrence matrix $\mathbf{M} = \{M_{i,j}\}$. The matrix item $P_{i,j}$ can be calculated as:

$$P_{i,j} = \frac{M_{i,j}}{n_i} \tag{9}$$

where n_i denotes the number of occurrences of e_i in the training set.

To avoid noise caused by the distribution of co-occurrences and faint emotion correlations, we constructed an improved correlation matrix $\mathbf{I} = \{I_{i,j}\}$ by thresholding the correlation matrix $\mathbf{P}$. Therefore, the improved correlation matrix can be formulated as:

$$I_{i,j} = \begin{cases} 0, & if \ \ p_{i,j} < \tau \\ 1, & if \ \ p_{i,j} \geq \tau \end{cases} \tag{10}$$

where τ is a threshold parameter to determine whether the correlation between two emotions is preserved. Based on $\mathbf{P}$ and the prior knowledge of the co-occurrence of positive emotions in movie clips, we observe that the inter-emotion correlation can be indicated when $P_{i,j}$ is greater than 0.4. We further choose $\tau = 0.4$, 0.5, and 0.6 for comparisons, and $\tau = 0.5$ leads to the optimal performance as shown in Table 1. Therefore, τ is set to 0.5 in our experiments.

TABLE 1
Results of *TransEEG* with $\tau = 0.4, 0.5, 0.6$ on the main metrics. "HL", "OE", "RL" are Hamming Loss, One-error, Ranking Loss respectively. "↓" indicates the smaller the better. "↑" indicates the larger the better.

	HL↓	OE↓	RL↓	mAP↑	Micro-F1↑
$\tau = 0.4$	0.360	0.720	0.278	0.510	0.600
$\tau = 0.5$	**0.356**	**0.719**	**0.277**	**0.513**	**0.619**
$\tau = 0.6$	0.360	**0.719**	0.279	0.510	0.603

Specifically, the element $I_{i,j} = 1/0$ denotes the presence/absence of correlation between e_i and e_j. Then the Inter-Emotion Mask can be calculated by:

$$M_{ER_{i,j}} = \begin{cases} 0, & if \;\; I_{i,j} = 1 \\ -\infty, & if \;\; I_{i,j} = 0 \end{cases} \tag{11}$$

where $\mathbf{M}_{ER} = \{M_{ER_{i,j}}\} \in \mathbb{R}^{l \times l}$.

Label-label correlation learning focuses on modeling the emotion dependencies via applying masked multi-head attention to a sequence of emotion word embeddings. Therefore, the informative emotion word embedding $\mathbf{E} \in \mathbb{R}^{l \times d_l}$ is used as input to the Transformer. Before being fed to the masked multi-head attention layer, the emotion embedding matrix $\mathbf{E}$ is projected to query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices by an unbiased linear projections. With the designed $\mathbf{M}_{ER}$, the self-attention scores in each "head" of masked multi-head attention is calculated by:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M}_{ER}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}_{ER}}{\sqrt{d_k}})\mathbf{V} \tag{12}$$

Under the guidance of Inter-Emotion Mask, the label correlation learning module outputs the label representation incorporating information about emotion dependencies, which is represented as $\mathbf{O}_E$. Therefore, the output of the multi-layer Transformer is projected via linear projection:

$$\mathbf{L} = linear(\mathbf{O}_E) \tag{13}$$

After that, $\mathbf{L}$ is utilized to generate the representation-label correlation matrix.

3.3.4 Correlation Guided Representation Learning Module

The EEG discriminative representation can be learned by leveraging both the spatial-temporal EEG features and label-label correlations. To refine the emotion-relevant EEG features, our model calculates a weighted sum of the EEG features by constructing an attention vector α:

$$\alpha = \tanh(\text{softmax}(\mathbf{G})), \mathbf{G} = \mathbf{R}\mathbf{L}^T \tag{14}$$

where $\mathbf{G} \in \mathbb{R}^{n \times l}$ is the representation-label correlation matrix, which is calculated by multiplying the spatial-temporal EEG features $\mathbf{R}$ with the label representation $\mathbf{L}$.

Correlation Guided Representation Learning generates the discriminative EEG representation $\mathbf{R}^{'} \in \mathbb{R}^{N \times d}$ by applying the attention score vector α to all sequences of spatial-temporal EEG features $\mathbf{R}$:

$$\mathbf{R}^{'} = \mathbf{R} \cdot \alpha \tag{15}$$

3.3.5 Classifier Learning

The multi-label positive emotion categories can be predicted by feeding the EEG representation $\mathbf{R}^{'}$ to the classifier, where we apply the sigmoid function to linear projections of the emotion-related EEG representation for classification:

$$p = \text{sigmoid}(\mathbf{R}^{'}\mathbf{W}_c + \mathbf{b}_c) \tag{16}$$

where $\mathbf{W}_c \in \mathbb{R}^{d \times l}$ and $\mathbf{b}_c \in \mathbb{R}^l$ are learnable parameters of the linear projection layer, and p is the positive emotion distribution. The sigmoid function is selected for multi-label classification as it can deal with non-exclusive labels.

To learn the mapping of multi-label positive emotion classification model from $\mathbf{R}^{'}$ to $2^{\mathcal{Y}}$, we use the binary cross-entropy loss function to measure probabilistic errors in the multi-label positive emotion classification task:

$$\mathcal{L}_{BCE} = -\frac{1}{l}\sum_{i=1}^{l}[y^i \log(p^i) + (1 - y^i)\log(1 - p^i)] \tag{17}$$

where y^i represents the ground truth label of e_i, and p^i denotes the predicted probability of the corresponding emotion.

3.4 Training Objective

Algorithm 1 summarizes the detailed training procedure of our *TransEEG* model. Although we use a five-fold cross-validation strategy to train the model, we will illustrate our method using a single split of the training set and test set for the sake of simplification. Denote $\mathcal{X} = [\mathcal{X}_S, \mathcal{X}_T]$ to be the entire data with $\mathcal{X}_S$ as a set of training data and $\mathcal{X}_T$ as a set of test data, and denote $\mathcal{Y}_S$ to be the ground truth label set associated with $\mathcal{X}_S$.

3.4.1 Implementation Details

We train our *TransEEG* on NVIDIA GTX TITAN X GPU, and the optimal hyperparameters are selected after extensive experimental comparisons as follows. The Adam optimizer is used to optimize our model with a batch size of 64. The maximum training epoch is set to 20. The learning rate is initially set to $1e$-3 and then decayed to 0.1 of the previous learning rate every 5 epochs according to a learning rate decay strategy. We also employed dropout in the linear layer, position encoder and Transformer encoder, and the dropout ratio is set to 0.5. The weight decay is set to $1e$-4 to prevent over fitting. Our spatial-temporal encoding module and label correlation learning module are designed based on a 2-layer Transformer encoder, with 6 "heads" in multi-head attention sub-layer and the hidden dimension of feed-forward network sub-layer is 1,024.

Algorithm 1 Training of *TransEEG*

Input:

1: Training data set $\mathcal{X}_S$ with the corresponding ground-truth label set $\mathcal{Y}_S$;
2: The channel sequence of EEG feature $\mathbf{X}^C$ and the temporal sequence of EEG feature $\mathbf{X}^S$;
3: The emotion word embedding sequence $\mathbf{E}$ and the emotion co-occurrence matrix $\mathbf{M}$;
4: The threshold τ for binarizing emotion correlation matrix $\mathbf{P}$;
5: The hyper-parameters, such as learning rate r, dropout p_d, weight decay w, etc.

Output:

The optimal emotion prediction probability vector p.

6: Initializing the model parameters $\theta = \{\theta_{ST}, \theta_{LC}\, \theta_{CR}\}$.
7: The spatial-temporal encoding module learns the spatial-temporal EEG features $\mathbf{R}$, and the parameters of the spatial-temporal encoding module can be represented as θ_{ST}.
8: Furthermore, the label correlation learning module focuses on modeling the label-label correlations, and outputs the label presentation $\mathbf{L}$. The parameters of the label correlation learning module can be represented as θ_{LC}.
9: Based on spatial-temporal EEG features $\mathbf{R}$ and label correlation representation $\mathbf{L}$, the correlation guided representation learning module aims to generate discriminative EEG representation $\mathbf{R}^{'}$ by calculating the attention score vector α from the representation-label correlation matrix $\mathbf{G}$.
10: Obtaining the emotion prediction probability vector p through Eq. 16.
11: Calculating the loss function by Eq. 17.
12: Using $\mathcal{X}_S$ and $\mathcal{Y}_S$ to update the parameters of *TransEEG* by gradient back-propagation:
$\theta_{ST} \leftarrow \theta_{ST} - r\frac{\partial \mathcal{L}_{BCE}}{\partial \theta_{ST}}$, $\theta_{LC} \leftarrow \theta_{LC} - r\frac{\partial \mathcal{L}_{BCE}}{\partial \theta_{LC}}$;
$\theta_{CR} \leftarrow \theta_{CR} - r\frac{\partial \mathcal{L}_{BCE}}{\partial \theta_{CR}}$;
13: Go to step 2, **until** the iterations satisfy the predefined algorithm convergence condition.

4 Experiments

To evaluate the effectiveness of our proposed method for multi-label positive emotion classification based on EEG signals, we conduct extensive experiments on a newly proposed positive emotion EEG database named CPED [18] with validated positive emotion-inducing movie clips.

4.1 Database and Protocols

CPED Database. The positive emotion EEG database consists of effectively validated emotion-inducing materials (movie clips) and biological signals (EEG signals). The CPED database consists of 22 movie clips under the 15 positive emotion categories. After clustering analysis, 9 movie clips with best induction effect under each main emotion category in CPED were selected for the present study, and the corresponding 9 positive emotions are the movie induced positive emotions analyzed in this paper. Furthermore, the 9 emotion categories (Tenderness, Gratitude, Romance, Amusement, Friendship, Awe, Craving, Respect, and Pride) are consistent with theoretical expectations [26] and are supported by existing research [16], [27]. And these discrete positive emotions have been considered to be the most commonly experienced emotions in daily life and could be identified based on EEG responses [16], [17], [18].

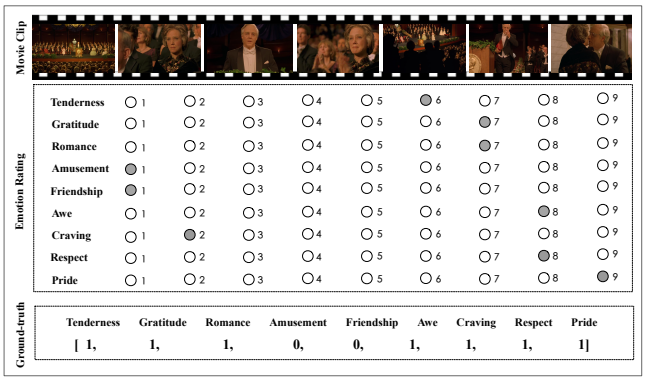

Fig. 5. Here, we use a clip from the movie *A Beautiful Mind* to show the rating score of emotion categories and the procedure of the ground-truth label annotation during participants watching the clip. Seven positive emotions were evoked (Since the rating score 1 means "not at all" according to [18], we assume that the corresponding emotion category does not appear in the video clip.) by this movie clip. The figure clearly shows the corresponding rating scores of emotion and the multi-label positive emotion ground-truth labels to the movie clip.

According to the international 10-20 system, EEG signals were recorded through 32 electrodes (two of which were reference electrodes) while 312 Chinese subjects watched 9 movie clips of 1-3 minutes in length. Afterwards, a bandpass filter with a frequency range of 1.0-45.0Hz was applied to EEG signals. We firstly decompose an EEG channel into 5 frequency bands, including the δ band (1-3Hz), the θ band (4-7Hz), the α band (8-13Hz), the β band (14-30Hz) and the γ band (31-45Hz). Then we extracted PSD features from each EEG channel using STFT with a 2-seconds sliding Hamming window with 50% overlap across the five sub-frequency bands. Therefore, the PSD dimension of five sub-frequency bands in a channel extracted through a sliding window is 5. Consequently, we can obtain the time-frequency domain features for each instance (where the time sequence length is 30, the number of channels is 30 and the feature dimension is 5).

Furthermore, PSD features from multi-channel EEG data are available for fulfilling our task. In our experiments, we train our method by minimizing the loss of multi-label classification. Therefore, we annotate the ground-truth emotion labels corresponding to the EEG data with respect to the participants' ratings. The procedure of label annotation is shown in Fig. 5. For each trial, participants rated emotional scores on a 9 point scale (1 = "not at all", 9 = "extremely" [18]). The scores represent the intensity of the emotions that appear in the corresponding movie clips. For instance, a participant rated an emotion category as 1, which indicates that this emotion category did not appear in the movie clip. For each EEG trial, an annotation of the EEG recording with 9 discrete positive emotion labels was assigned. Specifically, the ground-truth label annotation is a binary vector, where 0/1 indicates the absence/presence of

the corresponding emotion in the movie clip. Thus, based on the scoring of multiple labels, the ground-truth labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_i, \ldots, y_l\}$ and $y_i = 0/1$ are available, as shown in Fig. 5.

Experiment Protocols. Based on the CPED positive emotion database, we conducted experiments with a subject-independent five-fold cross-validation strategy to evaluate the performance of EEG-based multi-label positive emotion classification method. Specifically, the five-fold cross-validation is adopted to obtain stable and reliable models. In this paper, we follow the subject-independent principle to divide the data into five groups according to the number of subjects, and for each trial we select one of the five groups as the test data and the other groups as the training data. Furthermore, in order to investigate the impact of each module in our model, we conducted the ablation studies based on the variants of our proposed model.

4.2 Baseline and Evaluation Metrics

We compare our method with several representative multi-label learning algorithms [29]. 1) Problem transformation methods including first-order approach **Binary Relevance (BR)** [55] and high-order approach **Classifier Chains (CC)** [56] which transform the task of multi-label learning into the task of binary classification, second-order approach **Calibrated Label Ranking (CLR)** [57] which transforms the task of multi-label learning into the task of label ranking, and another high-order approach **Random k-labelsets (RKL)** [58] which transforms the task of multi-label learning into the task of multi-class classification. 2) Algorithm adaptation method **ML-kNN** [59], which tackles the multi-label classification problem by adapting popular machine learning techniques to deal with multi-label data directly.

The methods mentioned above have achieved remarkable performance in the field of multi-label classification and are mainly based on traditional machine learning methods. However, our proposed approach is based on deep models, and to be impartial, we adopt DBN [60] and DGCNN [20], which are two deep learning methods with significant performance in the field of EEG-based emotion recognition, as comparative models. The performance of all these methods are shown in Table 2.

To evaluate the performance of the multi-label classification task, we adopted eight widely used multi-label metrics after thorough consideration. Following the previous work [29], [61], [62], we employed **Hamming Loss**, **One-error**, **Ranking Loss**, **mAP** and **Micro-F1** as the main metrics, while the **Micro-accuracy**, **Micro-precision** and **Micro-recall** are also reported as reference. Additionally, we also report the **Macro**-based metrics to quantitatively assess the experimental performance from a different evaluation perspective. The detailed metric definitions can be found in [29]. For **Hamming loss**, **One-error** and **Ranking loss**, the smaller the values the better the performance (↓). For the other metrics, the bigger the values the better the performance (↑).

4.3 Experimental Results

4.3.1 Results of multi-label classification prediction

Table 2 shows the comparison results on CPED database, and the performance rank of each method comes after the measure values. Since eight measures are utilized in the experiments, the average ranks are shown at the bottom of each row.

Our *TransEEG* method presents the best performance of the main metrics. Compared to the traditional multi-label classification model, our *TransEEG* method decreases by 0.4% on hamming loss compared with Calibrated Label Ranking and improves by 10.9% on mAP over the Random k-labelsets. In addition, our *TransEEG* achieves 0.719 on OneError, which is 0.8% better than the best existing method of DGCNN. As for the comparison with deep learning methods DBN and DGCNN, *TransEEG* achieves the best performance in all evaluation metrics except the Micro-R. The DBN model achieves a result of 1.0 on the Micro-R evaluation metric, but *TransEEG* can only achieve 0.659. However, our *TransEEG* still ranks 2nd in terms of results in Micro-R. In experiments we expect both Micro-P and Micro-R to be as high as possible. Although DBN achieves the highest result of 1.0 in Micro-R, but the result in Micro-P was merely 0.434, while our model achieves superior results in both Micro-P and Micro-R metrics. However, even though Classifier Chain achieves the best performance amongst all traditional classification methods on the evaluation metrics Micro-F1 and Micro-P, it is worth noting that our *TransEEG* is no less impressive and achieves the 2nd ranking. We also provide the rank-based metrics in Table 2 to quantify the prediction performance, and our proposed *TransEEG* shows the best performance. Specifically, *TransEEG* decreases the Ranking Loss by 0.3% compared with DGCNN.

In summary, *TransEEG* performs the best overall, and achieves significant results compared to existing deep models which only learn the EEG representation without considering the correlation between labels. Therefore, to some extent, it shows the superiority of our model in discriminative EEG representation learning by capturing rich spatial-temporal information and exploiting label correlations.

Fig. 6 shows the Macro-based metrics to quantify the assessment of experimental performance from different evaluation perspectives. It is obvious that Our *TransEEG* achieves the best performance on mAP, so our method is more robust in multi-label positive emotion classification. In addition, *TransEEG* achieves superior results on Macro-A and Macro-R, respectively. We observe that the performance of our method on Macro-P and Macro-F1 is slightly inferior to DBN. As the Macro-P metric measures the average of precision over all categories, unbalanced data distribution can reduce the results. Therefore, our method is slightly influenced by the rare positive categories in the database according to experimental results. As shown in Fig. 7, the positive emotions are imbalanced in CPED, which is similar to the previous emotional databases. The Friendship, Romance and Amusement emotion categories account for a relatively small proportion of the database, and is approximately one-third of the amount of Respect emotion respectively. To address the data unbalance issue on the multi-label positive emotion classification task, we plan to expand the database in the future work.

4.3.2 Results of Ablation Studies

In order to understand the effect of each module in *TransEEG*, we report the results on CPED using the variant

TABLE 2
Comparison results of different multi-label classification methods on CPED. Note that Micro-A, Micro-P, Micro-R denote Micro-accuracy, Micro-precision, Micro-recall, respectively. "Average Rank" is calculated at the end of each column to demonstrate the overall performance, as each metric is a reflection on a certain aspect. "↓" indicates the smaller the better. "↑" indicates the larger the better.

Criterion	Traditional Models					Deep Models		Ours
	BR	**CC**	**CLR**	**RKL**	**ML-KNN**	**DBN**	**DGCNN**	**TransEEG**
Hamming Loss↓	0.360(2)	0.365(4)	0.360(2)	0.366(5)	0.372(6)	0.566(7)	0.364(3)	**0.356(1)**
OneError↓	0.822(4)	0.822(4)	0.822(4)	0.822(4)	0.821(3)	0.727(2)	0.727(2)	**0.719(1)**
Ranking Loss↓	0.532(5)	0.502(3)	0.532(5)	0.529(4)	0.577(6)	0.280(2)	0.280(2)	**0.277(1)**
mAP↑	0.399(4)	0.391(5)	0.399(4)	0.404(3)	0.345(6)	0.503(2)	0.503(2)	**0.513(1)**
Micro-F1↑	0.590(6)	**0.630(1)**	0.590(6)	0.602(4)	0.572(7)	0.605(3)	0.595(5)	0.619(2)
Micro-A↑	0.639(2)	0.634(4)	0.639(2)	0.633(5)	0.627(6)	0.434(7)	0.636(3)	**0.644(1)**
Micro-P↑	0.600(4)	**0.717(1)**	0.600(4)	0.643(3)	0.575(6)	0.434(7)	0.578(5)	0.644(2)
Micro-R↑	0.581(4)	0.562(7)	0.581(4)	0.569(6)	0.571(5)	**1.000(1)**	0.620(3)	0.659(2)
Average Rank	3.875(4)	3.645(3)	3.875(4)	4.250(5)	5.625(6)	3.875(4)	3.125(2)	**1.375(1)**

TABLE 3
Ablation study results of the variant models on CPED database. Note that Micro-A, Micro-P, Micro-R denote Micro-accuracy, Micro-precision, Micro-recall, respectively. The meanings of "↑", "↓" and "Average Rank" are the same as Table 2.

Criterion	The Variant Models				Ours
	LSTMEEG	**TransEEG**$_{\text{w/o Temporal}}$	**TransEEG**$_{\text{w/o Spatial}}$	**TransEEG**$_{\text{w/o LR}}$	**TransEEG**
Hamming Loss↓	0.372(4)	0.358(2)	0.360(3)	0.360(3)	**0.356(1)**
mAP↑	0.483(5)	0.509(3)	0.508(4)	0.510(2)	**0.513(1)**
Micro-F1↑	0.595(5)	0.604(4)	0.607(2)	0.606(3)	**0.619(1)**
Micro-A↑	0.621(5)	0.643(2)	0.641(3)	0.640(4)	**0.644(1)**
Micro-P↑	0.623(5)	0.643(2)	0.641(3)	0.640(4)	**0.644(1)**
Micro-R↑	0.600(5)	0.607(4)	0.627(2)	0.626(3)	**0.659(1)**
Average Rank	4.833(4)	2.833(2)	2.833(2)	3.167(3)	**1.000(1)**

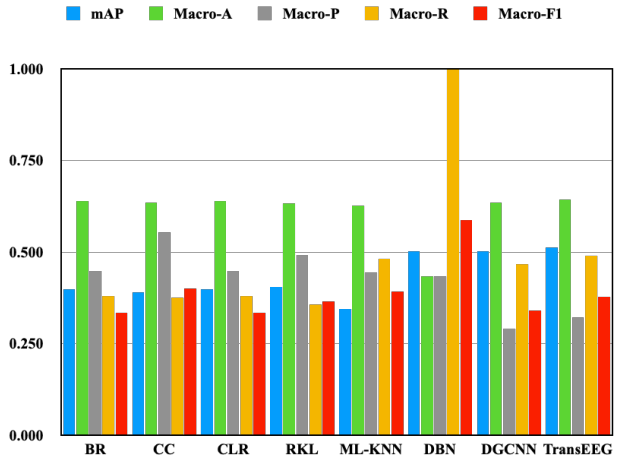

Fig. 6. The performance of different models on Mean Average Precision and Macro-based metrics. Note that mAP, Macro-A, Macro-P, Macro-R denote Mean Average Precision, Macro-Accuracy, Macro-precision, Macro-recall and Macro-F1, respectively.

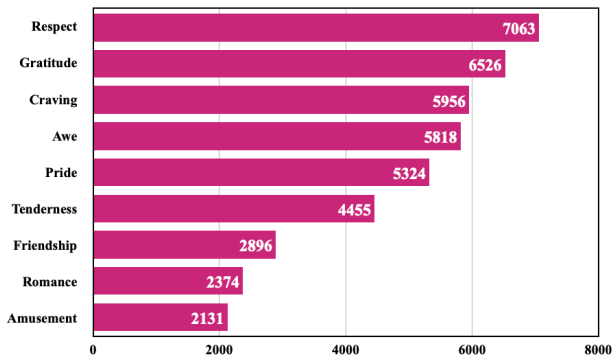

Fig. 7. Positive emotions distribution in CPED.

models by removing each module. In this section, we adopt **Hamming Loss**, **mAP**, **Micro-F1**, **Micro-A**, **Micro-P** and **Micro-R** as the evaluation metrics.

We evaluate our model from three perspectives. Firstly, to demonstrate whether the Transformer encoder module is superior to the LSTM module in terms of learning sequential dependencies, we replace the spatial encoder in **TransEEG**$_{\text{w/o Temporal}}$ with the LSTM for learning channel dependencies, and the variant model is denoted as **LSTMEEG**. Secondly, we evaluate the effectiveness of modeling long-distance dependencies among EEG channels and validate that temporal dependencies learning is effective in extracting discriminative features. Thus, we remove the spatial encoder and temporal encoder from our *TransEEG* separately, and get two variant models **TransEEG**$_{\text{w/o Spatial}}$ and **TransEEG**$_{\text{w/o Temporal}}$ from our *TransEEG* model. Thirdly, to evaluate the effect of label-label correlation in emotion-relevant EEG representation learning, we remove the label correlation learning module from our *TransEEG* and get the variant model **TransEEG**$_{\text{w/o LR}}$. We conduct experiments based on the above variant models, and the results of ablation studies are shown in Table 3.

(1) Transformer vs. LSTM: With our Transformer encoder modules, *TransEEG* has made a significant improvement on mAP metric compared to LSTMEEG method. **TransEEG**$_{\text{w/o Temporal}}$ extracts the spatial features from multi-channel EEG signals, and compared to LSTMEEG, the experimental results indicate that the ability of learning sequential dependent information is superior to LSTM.

(2) Spatial & Temporal modeling: From the overall results of ablation studies, we argue that characteristics of Transformer encoder to settle long-distance dependencies among EEG channels and between temporal sequences can assist in capturing more meaningful features. What's more, comparing the results of **TransEEG**$_{\mathrm{w/o\ Spatial}}$ and *TransEEG*, we can observe that *TransEEG* achieves a significant improvement on all the metrics. It shows that the information complementary of spatial correlation and the temporal dependencies in EEG signals are both essential and crucial for discriminative EEG representation learning.

(3) Label-label correlation: Without the module of label correlation learning, the correlation guided representation learning process in **TransEEG**$_{\mathrm{w/o\ LR}}$ is also omitted. The EEG representation for multi-label classification in **TransEEG**$_{\mathrm{w/o\ LR}}$ is the spatial-temporal encoding. Compared with the results of **TransEEG**$_{\mathrm{w/o\ LR}}$, *TransEEG* achieves the superior results, especially there is a 1.3% improvement on the Micro-F1 metric.

By comparing the results of ablation studies, our *TransEEG* model achieves the best performance over all the evaluation metrics.

5 Conclusion, Limitations and Future Work

In this paper, we propose a novel movie-induced multi-label positive emotion classification method named *TransEEG* through viewers' EEG signals. Our method exploits the correlation guided EEG representation for emotion classification by integrating the spatial-temporal EEG encoding with the label-label (i.e. inter-emotion) correlations. Specifically, we employed the Transformer encoder to extract the EEG spatial-temporal features and learn the inter-emotion correlations with the help of an Inter-Emotion Mask we designed. Comparative experiments demonstrate the superiority of our *TransEEG* method over the representative multi-label classification methods and the EEG-based emotion recognition neural models.

The perspective of analyzing movie induced multi-label positive emotions through the viewer's multi-channel EEG signals is a new task, which can be regarded as a successful attempt for the studies of movie induced positive emotion analysis in affective computing community. Due to the concomitant nature of positive emotion categories, the multi-label positive emotion analysis method is more suitable for movie-induced positive emotion recognition. Moreover, it is of great significance to utilize the inter-emotion correlations in the process of positive emotion analysis.

While undertaking this work, we have identified some limitations in our current study and found several directions for future improvement. Firstly, through the analysis of experimental results we found that the performance on Macro-based metrics is undesirable, which indicates that our method does not achieve ideal performance in identifying each emotion category exactly, due to the data imbalance issue. We will address the issue by expanding the EEG data to guarantee a balanced database. Meanwhile, we will attempt to adjust the loss function by penalizing the misclassified minority class, i.e., adding a penalty coefficient to the misclassified minority class samples to make the model more sensitive to them and thereby be able to identify the minority class samples more effectively.

Secondly, the EEG signals in CPED were recorded through 32 electrodes, according to the international 10-20 system. The EEG equipment employed demands a strict laboratory environment, so it limits the application of EEG-based positive emotion analysis in the real world. In recent years, the portable and stable EEG devices have been further developed, increasing the feasibility of conducting large-scale experiments in real-world applications. The number of electrodes in portable EEG devices is relatively small, so it is relevant for future work to explore the location of electrodes for efficient recognition of positive emotions. Meanwhile, improving the stability and scalability of our method enables to promote the popularity of applications.

Finally, there are many other factors that influence movie induced positive emotions, such as movie genres, movie aesthetic highlights, and viewer's personality. This will enable us to take multimodal information into account for investigating multi-label positive emotions in future work, not only in terms of viewers' physiological responses but also with respect to movie genres and movie aesthetics.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (4212029), Natural Science Foundation of China (62272447) and Newton Prize 2019 China Award (NP2PB/100047), and supported in part by Tsinghua University Initiative Scientific Research Program (20211080093) and the Natural Science Foundation of China (61725204).

References

[1] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "Movienet: A holistic dataset for movie understanding," in *European Conference on Computer Vision*. Springer, 2020, pp. 709–727.

[2] C. R. Plantinga and G. M. Smith, *Passionate views: Film, cognition, and emotion*. Johns Hopkins University Press, 1999.

[3] D. Zillmann, "Mood management: Using entertainment to full advantage," *Communication, Social Cognition, and Affect*, pp. 147–171, 2015.

[4] L. J. Rufer MD, "Magic at the movies: Positive psychology for children, adolescents and families," 2014.

[5] B. L. Fredrickson, M. A. Cohn, K. A. Coffey, J. Pek, and S. M. Finkel, "Open hearts build lives: positive emotions, induced through loving-kindness meditation, build consequential personal resources." *Journal of Personality and Social Psychology*, vol. 95, no. 5, p. 1045, 2008.

[6] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American Psychologist*, vol. 55, no. 1, p. 68, 2000.

[7] B. Campos, M. N. Shiota, D. Keltner, G. C. Gonzaga, and J. L. Goetz, "What is shared, what is different? core relational themes and expressive displays of eight positive emotions," *Cognition & Emotion*, vol. 27, no. 1, pp. 37–52, 2013.

[8] X. Hu, C. Zhuang, F. Wang, Y.-J. Liu, C.-H. Im, and D. Zhang, "fnirs evidence for recognizably different positive emotions," *Frontiers in Human Neuroscience*, vol. 13, p. 120, 2019.

[9] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[10] J. Wei, X. Yang, and Y. Dong, "User-generated video emotion recognition based on key frames," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 14 343–14 361, 2021.

[11] M.-C. Sun, S.-H. Hsu, M.-C. Yang, and J.-H. Chien, "Context-aware cascade attention-based rnn for video emotion recognition," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–6.

[12] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 143–10 152.

[13] M. N. Shiota, B. Campos, C. Oveis, M. J. Hertenstein, E. Simon-Thomas, and D. Keltner, "Beyond happiness: Building a science of discrete positive emotions." *American Psychologist*, vol. 72, no. 7, p. 617, 2017.

[14] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[15] M. N. Shiota, S. L. Neufeld, W. H. Yeung, S. E. Moser, and E. F. Perea, "Feeling good: autonomic nervous system responding in five positive emotions." *Emotion*, vol. 11, no. 6, p. 1368, 2011.

[16] X. Hu, J. Yu, M. Song, C. Yu, F. Wang, P. Sun, D. Wang, and D. Zhang, "Eeg correlates of ten positive emotions," *Frontiers in Human Neuroscience*, vol. 11, p. 26, 2017.

[17] G. Zhao, Y. Zhang, G. Zhang, D. Zhang, and Y.-J. Liu, "Multi-target positive emotion recognition from eeg signals," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020. [Online]. Available: doi:10.1109/TAFFC.2020.3043135

[18] Y. Zhang, G. Zhao, Y. Shu, Y. Ge, D. Zhang, Y.-J. Liu, and X. Sun, "Cped: A chinese positive emotion database for emotion elicitation and analysis," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021. [Online]. Available: doi: 10.1109/TAFFC.2021.3088523

[19] P. Keelawat, N. Thammasan, M. Numao, and B. Kijsirikul, "A comparative study of window size and channel arrangement on eeg-emotion recognition using deep cnn," *Sensors*, vol. 21, no. 5, p. 1678, 2021.

[20] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2018.

[21] X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu, and H. Wang, "An efficient lstm network for emotion recognition from multichannel eeg signals," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020. [Online]. Available: doi:10.1109/TAFFC.2020.3013711

[22] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2909–2917.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[24] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from eeg signals," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 550–562, 2017.

[25] L. F. Barrett, J. Gross, T. C. Christensen, and M. Benvenuto, "Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation," *Cognition & Emotion*, vol. 15, no. 6, pp. 713–724, 2001.

[26] B. L. Fredrickson, "Positive emotions broaden and build," in *Advances in Experimental Social Psychology*. Elsevier, 2013, vol. 47, pp. 1–53.

[27] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[28] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of Personality and Social Psychology*, vol. 54, no. 6, p. 1063, 1988.

[29] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.

[30] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal multi-label emotion detection with modality and label dependence," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3584–3593.

[31] B. Kostiuk, Y. M. Costa, A. S. Britto, X. Hu, and C. N. Silla, "Multi-label emotion classification in music videos using ensembles of audio and video features," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2019, pp. 517–523.

[32] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Latent emotion memory for multi-label emotion classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7692–7699.

[33] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 884–906, 2019.

[34] D. Zhang, X. Ju, W. Zhang, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing," in *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021, pp. 14 338–14 346.

[35] C. E. Waugh, "The roles of positive emotion in the regulation of emotional responses to negative events." *Emotion*, vol. 20, no. 1, p. 54, 2020.

[36] Z. Jin, Y. Yao, Y. Ma, and M. Xu, "Thuhcsi in mediaeval 2017 emotional impact of movies task." *MediaEval*, vol. 17, pp. 13–17, 2017.

[37] K.-A. C. Quan, V.-T. Nguyen, and M.-T. Tran, "Frame-based evaluation with deep features to predict emotional impact of movies." *MediaEval*, vol. 1, no. 2, p. 7, 2018.

[38] A. C. Micu and J. T. Plummer, "Measurable emotions: How television ads really work: Patterns of reactions to commercials can demonstrate advertising effectiveness," *Journal of Advertising Research*, vol. 50, no. 2, pp. 137–153, 2010.

[39] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[40] F. Silveira, B. Eriksson, A. Sheth, and A. Sheppard, "Predicting audience responses to movie content from electro-dermal activity signals," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2013, pp. 707–716.

[41] K. Kimura, S. Haramizu, K. Sanada, and A. Oshida, "Emotional state of being moved elicited by films: A comparison with several positive emotions," *Frontiers in Psychology*, vol. 10, p. 1935, 2019.

[42] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[43] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual lstm network," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 176–183.

[44] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "Sparsedgcnn: Recognizing emotion from multichannel eeg signals," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[45] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, 2019, pp. 4171–4186.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[47] 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 Journal*, vol. 22, no. 5, pp. 4359–4368, 2022.

[48] J.-Y. Guo, Q. Cai, J.-P. An, P.-Y. Chen, C. Ma, J.-H. Wan, and Z.-K. Gao, "A transformer based neural network for emotion recognition and visualizations of crucial eeg channels," *Physica A: Statistical Mechanics and its Applications*, vol. 603, p. 127700, 2022.

[49] J. Sun, J. Xie, and H. Zhou, "Eeg classification with transformer-based models," in *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, 2021, pp. 92–93.

[50] H. Zhu, F. Nan, Z. Wang, R. Nallapati, and B. Xiang, "Who did they respond to? conversation structure modeling using masked hierarchical transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9741–9748.

[51] J. Li, Z. Lin, P. Fu, Q. Si, and W. Wang, "A hierarchical transformer with speaker modeling for emotion recognition in conversation," *arXiv preprint arXiv:2012.14781*, vol. abs/2012.14781, 2020.

[52] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[53] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019.
[54] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1243–1252.
[55] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.
[56] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, p. 333, 2011.
[57] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multi-label classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, 2008.
[58] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.
[59] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.
[60] W. Zheng and B. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.
[61] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: sequence generation model for multi-label classification," in *Proceedings of the 27th International Conference on Computational Linguistics, COLING*, 2018, pp. 3915–3926.
[62] Q. Zhang, X. Zhang, Z. Yan, R. Liu, Y. Cao, and M. Zhang, "Correlation-guided representation for multi-label text classification," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, 2021, pp. 3363–3369.

Xiaobing Du is currently pursuing the Ph.D. degree in the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China. She received her bachelor degree from the School of software, Shandong University, Shandong, China, in 2016. Her research interests include affective computing and human-computer interaction.

Xiaoming Deng is currently a Professor with the Institute of Software, Chinese Academy of Sciences (CAS). He received the Bachelor and Master degrees from Wuhan University, and the Ph.D. degree from the Institute of Automation, CAS. He has been a Research Fellow at the National University of Singapore, and a Postdoctoral Fellow at the Institute of Computing Technology, CAS, respectively. His main research topics are in computer vision, and specifically related to 3D reconstruction, human motion tracking and synthesis, and natural user interfaces.

Hangyu Qin is currently a master student in University of Chinese Academy of Sciences, Beijing, China. She received her bachelor degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2020. Her research interests include affective computing and human-computer interaction.

Yezhi Shu is a Ph.D student with Department of Computer Science and Technology, Tsinghua University. She received her B.Eng. degree from Shandong University, China, in 2019. Her research interests include computer vision, deep learning algorithms and applications.

Fang Liu received her Ph.D. degree from the University of the Chinese Academy of Sciences (UCAS), Beijing, China, in 2021. She is currently a postdoc at Tsinghua University. Her research interests include computer vision, sketch interaction, and affective computing.

Guozhen Zhao received the B.S. degree in industrial engineering from Tianjin University, Tianjin, China, in 2007, and the M.S. and Ph.D. degrees in industrial and systems engineering from the State University of New York, Buffalo, NY, USA, in 2009 and 2011, respectively. Since 2012, he is now an Associate Professor with the Institute of Psychology, Chinese Academy of Sciences, Beijing, China. His current research interests include mathematical modeling of human cognition and performance, transportation safety, human computer interaction, and neuroergonomics and their applications in intelligent system design.

Yu-Kun Lai received his Bachelor and Ph.D. degrees in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a Professor in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial boards of *Computer Graphics Forum* and *The Visual Computer*.

Cui-Xia Ma received the B.S. and M.S. degrees from Shandong University, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Software, Chinese Academy of Sciences, Beijing, China, in 2003. She was a Research Associate in the Department of Computer Science, Naval Postgraduate School in Monterey, CA, USA, from 2005 to 2006. She is now a Professor with the Institute of Software, Chinese Academy of Sciences. Her research interests include human computer interaction and multimedia computing.

Yong-Jin Liu received the B.Eng. degree from Tianjin University, Tianjin, China, in 1998, and the M.Phil. and Ph.D. degrees from the Hong Kong University of Science and Technology, Hong Kong, China, in 2000 and 2004, respectively. He is now a Professor with Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include computational geometry, computer vision, cognitive computation and pattern analysis. For more information, visit http://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.htm

Hong-An Wang received the Ph.D. degree from the Institute of Software, Chinese Academy of Sciences, Beijing, China, in 1999. He is a Professor with the Institute of Software, Chinese Academy of Sciences. He is currently the Director of Intelligence Engineering Laboratory. His research interests include human-computer interaction, real-time intelligence, and real-time active database.