

Towards emotion recognition for virtual environments: an evaluation of eeg features on benchmark dataset

M. L. R. Menezes¹  · A. Samara² · L. Galway² · A. Sant'Anna¹ · A. Verikas¹ · F. Alonso-Fernandez¹ · H. Wang² · R. Bond²

Received: 22 April 2017 / Accepted: 22 July 2017
© The Author(s) 2017. This article is an open access publication

Abstract One of the challenges in virtual environments is the difficulty users have in interacting with these increasingly complex systems. Ultimately, endowing machines with the ability to perceive users emotions will enable a more intuitive and reliable interaction. Consequently, using the electroencephalogram as a bio-signal sensor, the affective state of a user can be modelled and subsequently utilised in order to achieve a system that can recognise and react to the user's emotions. This paper investigates features extracted from electroencephalogram signals for the purpose of affective state modelling based on Russell's Circumplex Model. Investigations are presented that aim to provide the foundation for future work in modelling user affect to enhance interaction experience in virtual environments. The DEAP dataset was used within this work, along with a Support Vector Machine and Random Forest, which yielded reasonable classification accuracies for *Valence* and *Arousal* using feature vectors based on statistical measurements and band power from the α , β , δ , and θ waves and High Order Crossing of the EEG signal.

Keywords Affective computing · Virtual environment · EEG · Emotion recognition · Feature extraction

This work was funded by the Science Without Borders program from the Brazilian government and EU COST Action TD1405.

✉ M. L. R. Menezes
maria.menezes@hh.se

¹ Center for Applied Intelligent Systems Research, Halmstad University, Halmstad, Sweden

² School of Computing and Mathematics, Ulster University Belfast, Belfast, UK

1 Introduction

Due to their increasing complexity, one of the main challenges found in virtual environments (VEs) is user interaction. Therefore, it is important to structure interaction modalities based on the requirements of the application, which may include both traditional and natural user interfaces, situational awareness and adaptation, personalised content management, multimodal dialogue and multimedia applications.

VEs typically require personalised interaction in order to maintain user engagement with the underlying task. While task engagement encompasses both the user's cognitive activity and motivation, it also requires an understanding of affective change in the user. Accordingly, physiological computing systems may be utilised to provide insight into the cognitive and affective processes associated with task engagement [15]. In particular, an indication of the levels of brain activity, through acquisition and processing of electroencephalogram (EEG) signals, may yield benefits when incorporated as an additional input modality [48].

In recent studies, EEG has been used to map the responses to the environment directly to the user's brain activity [1, 28, 33, 43, 49]. These systems are typically used for control purposes, enhancing traditional modalities such as mouse, keyboard, or game controller. However, this form of active interaction is still quite costly for users as it requires training and a good amount of both concentration and effort to modulate one's brain activity. This ultimately causes the user to focus more on the interaction modality itself than the underlying task. In order to achieve truly transparent interaction, the system is required to acquiesce to the user's intentions or needs. Using EEG as a bio-signal sensor to model the user's cognitive and affective state is

one potential way to achieve an interaction that does not require any training or attention focus from the user.

Many authors have investigated the use of EEG for recognizing user affect. However, EEG signals are complex, multi-modal time series and there is no consensus on which features are better suited for this task. The main contributions of this paper are twofold: (1) a summary of how affect recognition can augment VR environments targeting different applications, namely, medicine, education, entertainment and lifestyle; (2) an evaluation of several types of features for affect recognition using EEG on a benchmark dataset. For the purposes of the investigations, the DEAP dataset was used to provide an annotated set of EEG signals [24]. Support Vector Machine (SVM) and Random Forest were employed to classify different affective states according to the Circumplex Model.

2 Background

A system that can detect and adapt to user's current affective state is interesting for a broad range of applications, from medicine and education to entertainment, games and lifestyle.

2.1 Applications in medicine

VEs have been shown to help in the treatment of many conditions, as well as help people cope with distressing emotions such as anxiety and stress. Virtual Reality Exposure Therapy (VRET), for example, is an increasingly common treatment for anxiety and specific phobias [36]. When a user is immersed in a VE, they can be systematically exposed to specific feared stimuli within a contextually relevant setting [4, 6, 16, 17]. VEs have also been shown to help children with Autism Spectrum Disorders (ASD) improve their social functioning [3]. These examples indicate where a system that uses emotional modulation could be useful: to help the physician analyse the emotional states and development of the patient's condition, as well as to use that information to adapt the treatment in real-time, avoiding possible over exposure of the patient.

2.2 Applications in education

The association between Affective Computing and learning is known as Affective Learning (AL): technologies that sense and respond to affective states during the learning process to make knowledge transfer and development more effective[41]. The recognition that interest and active participation are important factors in the learning process are largely based on intuition and generalization of constructivist theories [7, 41]. AL can change this scenario by measuring, modelling, studying and supporting the affective

dimension of learning in ways that were not previously possible. Previous works have shown that VEs and AL can improve student performance [19, 27]. However, many of the previous approaches rely on questionnaires and other forms of off-line evaluation of affective state. The use of bio-sensors such as EEG might enable educational systems to automatically recognise affect and better understand non-verbal clues just as a teacher would.

2.3 Applications in entertainment and lifestyle

The entertainment industry is very enthusiastic regarding VEs, games being perhaps the most noticeable application. This enthusiasm is not surprising, to some degree, emotional experiences are what game designers create and sell [35]. Not only can VEs be designed to elicit both positive and negative emotions [13, 42], but also previous works have shown that emotion positively correlates with presence—the psychological sense of being in or existing in the VE in which one is immersed [2]. Another well-known use of VE in games are virtual worlds, such as Second Life [26]. The High Fidelity platform is able to track facial expressions in real time and transfer those to the user's avatar. Despite being able to mimic facial expressions related to speech and emotions, the system itself does not attempt to recognize affect [34]. EEG could extend the high fidelity platform with the ability to adapt to users' affect. It would also enable users that are unable to change their facial expressions—due to paralysis for example—to take advantage of a platform like High Fidelity. The use of the automatic modulation of user's emotional states in VEs are limitless and benefit from the proven relation between presence and emotional state.

2.4 EEG as an input modality for emotion recognition

Currently, various input modalities exist that can be utilised to acquire information about users and their emotions. More commonly, audiovisual-based communication, such as eye gaze tracking, facial expressions, body movement detection, and speech and auditory analysis may be employed as input modalities. Furthermore, physiological measurements using sensor-based input signals, such as EEG, galvanic skin response, and electrocardiogram can also be utilised. However, the use of EEG as an input modality has a number of advantages that make it potentially suitable for use in real-life tasks including its non-invasive nature and relative tolerance to movement. EEG can be used as a standalone modality as well as combined to other biometric sensors. The company *iMotions* for example has successfully developed a commercial platform for monitoring physiological and psychological parameters of users while experiencing VR. This is a great example of how affect recognition can be used to add value to VR applications [18, 20].

Several existing studies have exploited EEG as an input modality for the purpose of emotion recognition. Picard et al. looked at different techniques for feature extraction and selection in order to enhance emotion recognition by employing different biosignal data [40]. They found that there is a variation in physiological signals of the same subject expressing the same emotion from day to day. Which impairs recognition accuracy if not managed properly. Konstantinidis et al. studied real-time classification of emotions by analysing EEG data recorded using 19 channels. They showed that extracting features from EEG data using a complex non-linear computation, which is a multi-channel correlation dimension, and processing the features using a parallel computing platform (i.e. CUDA) would substantially reduce the processing time needed. Their method facilitates real-time emotion recognition [25].

Petrantonakis et al. proposes feature extraction methods based on Higher Order Crossing (HOC) analysis to recognise emotions from EEG data additionally to four different classification techniques. The highest reported classification accuracy was 83.33% using SVM trained on extracted HOC feature [37]. Murugappan investigated feature extraction using wavelet transforms [30]. Moreover, they used K-Nearest Neighbor to evaluate classification accuracy for emotions across two different sets of EEG channels (24 and 64 channels), with a resulting classification accuracy of 82.87%. Jenke et al. looked for feature selection methods extracted from EEG for emotion recognition [21]. They presented a systematic comparison of the wide range of available feature extraction methods using machine learning techniques for feature selection. Multivariate feature selection techniques performed slightly better than univariate methods, generally requiring less than 100 features on average.

Still there are challenges encountered when attempting to exploit EEG for emotional state recognition. Extracting relevant and informative features from EEG signals from a large number of subjects and formulating a suitable representation of this data in order to distinguish different affective states is an extremely complicated process [45]. This work utilizes a fairly large dataset of EEG signals to investigate the relevance of different features for the dimensions of Valence and Arousal, according to Russel's Circumplex Model of Affection. In this context, we aim to provide foundations for modelling user affect in order to enhance interaction experience in VEs.

3 Methodology

3.1 The DEAP dataset

The DEAP dataset [24], utilised in the work presented herein, comprises EEG and peripheral physiological signals

for 32 subjects who individually watched 40 one-minute music videos of different genres as a stimulus to induce different affective states. Within the dataset 32 channels were used to record EEG signals for each trial per subject, resulting in 8064 samples that represent the signal over each one-minute trial. During each trial, a single subject rated his/her feelings after watching the video using the Self Assessment Manikin (SAM) scale in the range [1–9] to indicate the associated levels of Valence, Arousal, Dominance and Liking

The DEAP is a benchmark dataset for emotion analysis using EEG, physiological and video signals developed by researcher of the Queen Mary University of London, United Kingdom; the University of Twente, The Netherlands; the University of Geneva, Switzerland; and the École polytechnique Fédérale de Lausanne, Switzerland. Even though it does not represent data used in VEs per se, its data is considered consistent by more than 560 citations from the research community and a good source for affective data in general.

3.2 Selection of EEG channels

Psycho-physiological research has shown that left and right frontal lobes have significant activity during the experience of emotions [32]. There is also evidence of the role of the prefrontal cortex in affective reactions and particularly in emotion regulation and conscious experiment [12]. Many scientific experiments have successfully used electrodes located in those regions to analyse affective states [10, 37].

Since the purpose of this work is to model user affect aiming real time applications, a simpler and more user-friendly environment for data acquisition is required. In an effort to reduce the number of electrodes, the signals were selected from four positions Fp1, Fp2, F3 and F4 only, according to the 10–20 system, as seen in Fig. 1.

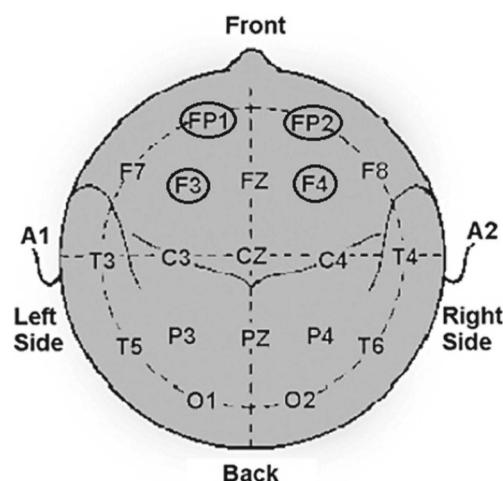


Fig. 1 Fp1, Fp2, F3 and F4 positions selected according to the 10–20 system [31]

3.3 Bandwave extraction

Commonly, brainwaves are categorized into four different frequency bands: Delta (δ) from 0.5 to 4 Hz; Theta (θ) from 4 to 8 Hz; Alpha (α) from 8 to 12 Hz; and Beta (β) 12 to 30 Hz. Literature has shown a strong correlation between these waves and different affective states [29].

The EEG data associated with each of the selected channels was transformed into α , β , δ , and θ waves, using the Parks–McClellan algorithm and Chebyshev Finite Impulse Response filter was applied to the signal according to the frequency ranges of each brainwave.

3.4 Feature extraction

Three types of features were computed from the EEG signal: statistical, powerband and High Order Crossing (HOC). Features along with the construction of the relevant feature vectors (FVS) are further explained within the following.

3.4.1 Statistical features

We adopted six descriptive statistics, as suggested by Picard et al. in [40] and Petrantonakis in [38]. The statistical features were extracted from the EEG signal in time domain and from each of the brainwaves, creating separated feature vector for both time and frequency domain:

- (a) Mean (μ)
- (b) Standard deviation (σ)
- (c) Mean of the absolute values of the first differences (AFD)
- (d) Mean of the normalised absolute values of the first differences (\overline{AFD})
- (e) Mean of the absolute values of the second differences (ASD)
- (f) Mean of the normalised absolute values of the second differences (\overline{ASD})

3.4.2 Spectral power density of brain waves

For the selected four channels, the mean log-transformed brain wave power were extracted from the α , β , δ , and θ frequency bands, according to [11]. The Spectral Power Density (SPD) is widely used to detect the activity level in each brain wave, allowing the components in the frequency domain to be interpreted as electroencephalographic rhythms.

For each electrode was calculated four features, representing the medium power of the signal for the entire bandwave, result in a 16-feature vector:

$$FV_{SPD} = [f_{F_p1}, f_{F_p2}, f_{F_3}, f_{F_4}]$$

Being each channel feature ($f_{F_{ch}}$) a feature vector of the mean power of the signal for the respective bandwave:

$$f_{ch} = [SPD_\alpha, SPD_\beta, SPD_\delta, SPD_\theta]$$

3.4.3 Higher order crossing

In this technique, a finite zero-mean time series $\{Z_t\}, t = 1, \dots, N$ oscillating through level zero can be expressed by the number of zero crossings (NZC). Applying a filter to the time series generally changes its oscillation and consequently its NZC. When a specific sequence of filters is applied to a time series, a specific corresponding sequence of NZC is obtained. This is called a High Order Crossing (HOC) sequence [22, 38].

The difference operator (∇) is a high-pass filter defined as $\nabla Z_t \equiv Z_t - Z_{t-1}$. A sequence of filters $\mathfrak{J}_k \equiv \nabla_{k-1}, k = 1, 2, 3, \dots$; and its corresponding HOC sequence, can then be defined as

$$D_k = NZC \{\mathfrak{J}_k(Z_t)\}, k = 1, 2, 3, \dots; t = 1, \dots, N$$

with

$$\mathfrak{J}(Z_t) = \nabla^{k-1} Z_t = \sum_{j=1}^k \frac{(k-1)!}{(j-1)!(k-j)!} (-1)^{j-1} Z_{t-j+1}$$

To calculate the number of zero-crossings, a binary time series is initially constructed given by:

$$X_t(k) = \begin{cases} 1, & \mathfrak{J}_k(Z_t) \geq 0 \\ 0, & \mathfrak{J}_k(Z_t) < 0 \end{cases}, k = 1, 2, 3, \dots; t = 1, \dots, N$$

Finally, the HOC sequence is estimated by counting the symbol changes in $X_1(k), \dots, X_N(k)$:

$$D_k = \sum_{t=2}^N [X_t(k) - X_{t-1}(k)]^2$$

In this paper filters up to order six were used, creating the feature vector $FV_{HOC} = [D_1, D_2, \dots, D_6]$.

3.5 Affective state classification

The Circumplex Model of Affect developed by James Russell suggests that the core of emotional states are distributed in a two-dimensional circular space, containing Arousal and Valence dimensions. Arousal is represented by the vertical axis and Valence is represented by the horizontal axis, while the center of the circle represents a neutral level of Valence and Arousal [44], as seen in Fig. 2.

As the current study is interested in recognising the affective state that a subject is experiencing, congruous with the two-dimensional Russell's Circumplex Model, throughout the investigations only Valence and Arousal ratings were used. Valence and Arousal ratings are provided within the DEAP dataset as numeric values ranging from [1–9] based on the SAM scale [5]. Two different partitioning schemes

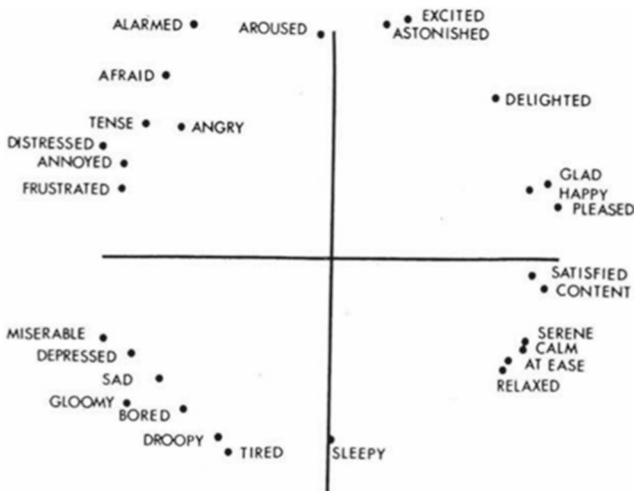


Fig. 2 Russel's Circumplex Model of Affect [44]

have been employed in order to discretize the range of values within the scale, as illustrated in Fig. 3, and given as follows:

- (a) *Tripartition Labeling Scheme*: Dividing the scale into three ranges [1.0–3.0], [4.0–6.0] and [7.0–9.0], given as the partitions *Low*, *Medium* and *High* respectively.
- (b) *Bipartition Labeling Scheme*: Similar to the previous scheme, however we removed instances annotated as *Medium*, resulting in the two ranges [1.0–3.0] and [7.0–9.0], given as the partitions *Low* and *High* respectively.

Within the research literature, a range of classification techniques have been used for affective computing and emotion recognition using EEG bio-signals as an input modality [23]. For the investigations presented herein we utilised two different classification methods: C-Support Vector Classification (SVM) with a linear kernel and Random Forest. The chosen SVM implementation is available from the LIBSVM library developed at National Taiwan University [9, 14] and the Random Forest developed by Leo Breiman [8].

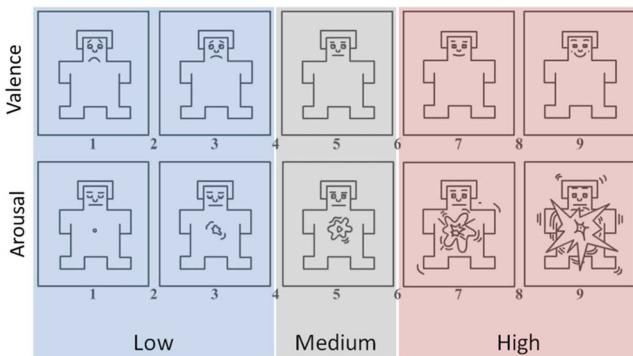


Fig. 3 Mapping from SAM scale Valence and Arousal values to Labels (*Low*, *Medium*, *High*)

Support Vector Machine (SVM) and Random Forest (RF) are versatile and widely used methods that have been shown to perform well in many application areas. The success of SVMs have been attributed to three main reasons: “their ability to learn well with only a small number of free parameters; their robustness against several types of model violations and outliers; and their computational efficiency compared to other methods.” [46]. Compared to other machine learning methods, RF present three interesting additional features: “a built-in performance assessment; a measure of relative importance of descriptors; and a measure of compound similarity that is weighted by the relative importance of descriptors” [47].

4 Experimental results

For the sake of exploration of different features, as previously described, we used classification accuracy as a metric. Furthermore, we have utilised the 10-fold cross validation approach for assessing classification performance. As previously discussed, this investigation aims to identify patterns related to features extracted from EEG signals across different Valence and Arousal states. For that, we applied SVM and Random Forest. Moreover, two labeling schemes were employed for each of the affective dimensions, i.e. *Bipartition* and *Tripartition*.

The following tables show the average results obtained for all the instances in the dataset, i.e. all videos for all participants. A comparison of the SVM and Random Forest results for all methods can be seen in Tables 1 and 2.

We can see that the results obtained for Random Forest were slightly better than SVM for all methods except Spectral Power Density. The comparison of the two tables show that the features extracted from the EEG signal behave similarly for any of the classification methods applied. Being the biggest difference for Statistic features extracted from Brainwaves for Arousal, that for Random Forest had 74.0% accuracy and for SVM only 57.2%. We can conclude that Random Forest performed better for all the Features in general and specially for Statistics of Brainwaves. SVM can be

Table 1 Classification accuracy per method, using SVM

Method	Bipartition		Tripartition	
	Arousal	Valence	Arousal	Valence
Statistics—Time	65.0%	61.2%	57.0%	51.3%
Statistics—Bandwaves	57.2%	83.2%	59.7%	55.1%
Bandwaves SPD	69.2%	88.4%	59.5%	55.9%
HOC	56.8%	62.7%	59.1%	53.5%

Table 2 Classification accuracy per method, using Random Forest

Method	Bipartition		Tripartition	
	Arousal	Valence	Arousal	Valence
Statistics—Time	67.1%	61.3%	57.7%	50.0%
Statistics—Bandwaves	74.0%	88.4%	63.1%	58.8%
Bandwaves SPD	67.9%	86.6%	56.1%	55.2%
HOC	57.4%	64.7%	57.8%	55.1%

a better choice if the chosen features are the spectral power density of Brainwaves and the class in interest is Valence.

Tables 1 and 2 show that Bipartition overcomes Tripartition for all methods tested except Arousal for HOC. Although the approximately 2% for Bi and Tripartition do not represent a statistically significant difference in accuracy. The best result for Tripartition is 63.1% for the Statistic features of the Brainwaves and Arousal in Table 2. Despite the results for Arousal in Tripartition being slightly better than the ones for Valence, the difference is not statistically significant.

The results are more interesting for Bipartition, in which the features tested are generally better representatives for Valence than Arousal, with an average difference of approximately 9% and the highest difference of approximately 18% for SPD.

We can also note that the best results were obtain for the methods that involve Bandwaves' features: Statistics and SPD. Valence has the best accuracies of 88.4 and 86.6%, respectively. The result for Arousal are 74.0 and 67.9% in Table 2.

Table 3 Classification accuracy for SPD using SVM

SPD	Bipartition		Tripartition	
	Arousal	Valence	Arousal	Valence
α	52.6%	73.1%	59.0%	57.3%
β	64.6%	69.8%	59.0%	55.6%
δ	66.2%	82.9%	60.2%	54.9%
θ	62.9%	76.1%	59.4%	55.9%
α, β	65.6%	82.7%	59.8%	56.1%
α, δ	66.5%	88.1%	59.4%	57.7%
α, θ	62.1%	83.4%	58.8%	55.1%
β, δ	66.9%	88.4%	59.1%	56.3%
β, θ	67.5%	85.4%	59.7%	55.6%
δ, θ	67.1%	88.9%	59.8%	55.2%
α, β, δ	67.5%	88.6%	59.9%	57.2%
α, β, θ	65.8%	85.0%	59.5%	56.8%
α, δ, θ	66.2%	88.7%	59.4%	57.0%
β, δ, θ	67.7%	88.4%	59.2%	54.2%
$\alpha, \beta, \delta, \theta$	69.2%	88.4%	59.5%	55.9%

Table 4 Classification accuracy for all Statistical Features for each Brainwave using Random Forest

Statistics—Bandwaves	Bipartition		Tripartition	
	Arousal	Valence	Arousal	Valence
α	66.6%	78.3%	59.7%	54.0%
β	67.5%	77.1%	59.3%	54.2%
δ	75.7%	87.9%	60.9%	55.3%
θ	71.3%	84.9%	61.3%	53.5%

The subsequent tables show the percentage of correctly classified instances for the methods that showed the best results: SPD using SVM, in Table 3; and Statistics of Brainwave using Random Forest, in Tables 4 and 5.

In Table 3 is clear that SPD features best relate to Valence in Bipartition, being δ 's SPD the best single feature with 82.9% accuracy. Combining two other features, such as the SPD of α and β or α and θ or even β and θ we can obtain similar results as δ alone: 82.7, 83.4 and 85.4%, respectively. Combining any of the single features with δ 's SPD increases the accuracy approximately 5%. The second best single feature is θ 's SPD and combining both δ 's and θ 's SPD gives the best result of 88.9%, better than combining all features in one single vector, 88.4%.

Table 4 shows the accuracy obtained for the Statistic features of each of the single Brainwaves using Random Forest. Being δ and θ again the brainwaves which features have the best results, 87.9 and 84.9% for Valence and 75.7 and 71.3% for Arousal in Bipartition. Combining the statistical features of the two bandwaves δ and θ increases the accuracy for Valence to 88.2%, almost the same as using the features for all bandwaves, 88.4%. Combining those same features for Arousal, on the other hand, gives the accuracy of 73.8%, worse than the result for δ only.

Table 5 shows the accuracy obtained for each of the single Statistic features for all Brainwaves combined using Random Forest. Here, we can see again the best results for Valence in Bipartition. For the single statistical features,

Table 5 Classification accuracy for all Brainwaves for each Statistical feature, using Random Forest

Statistics—Bandwaves	Bipartition		Tripartition	
	Arousal	Valence	Arousal	Valence
μ	64.1%	80.7%	55.9%	53.5%
σ	70.7%	87.4%	56.9%	54.8%
AFD	64.8%	89.9%	54.8%	54.9%
\overline{AFD}	68.3%	71.3%	58.4%	53.5%
ASD	67.7%	88.4%	57.7%	55.5%
\overline{ASD}	71.7%	73.8%	57.9%	53.8%

AFD has the best result of 89.9%, followed by ASD and σ , with 88.4 and 87.4%, respectively. Combining the three set of features again does not give a better accuracy than the best single feature, resulting in 88.6%.

For Valence, on the other hand, the best features are \overline{ASD} , σ , \overline{AFD} and ASD , with 71.7, 70.7, 68.3 and 67.7% classification accuracy, respectively. Combining those features does not improve the accuracy, resulting in 68.8% classification accuracy.

5 Discussion

The investigations and associated results presented in this paper show the potential of utilizing EEG signal data for recognising affective states. Based on the classification accuracy, the approach could be used to effectively recognise emotions in certain types of virtual reality environments. Educational applications could benefit from it by adapting the content of a course to the students anxiety levels, characterised by low levels of arousal and valence, detected by the bipartition approach. Other than that, the approach presented could be applied to medical applications that aim to help patients deal with phobias or entertainment platforms for social anxiety.

Both classification methods applied gave similar results, being the results for Random Forest slightly better than the ones for SVM. Particularly, the highest classification accuracy was obtained using the feature vector generated based on the statistical measurements derived from brainwaves, e.g 88.4% for Valence and 74% for Arousal.

Likewise, using a feature vector based on the associated power bands and SVM also produced the classification accuracy of 88.4% for Valence and slightly lower for Arousal, 69.2%. In both cases, the *Bipartition* labelling scheme was used.

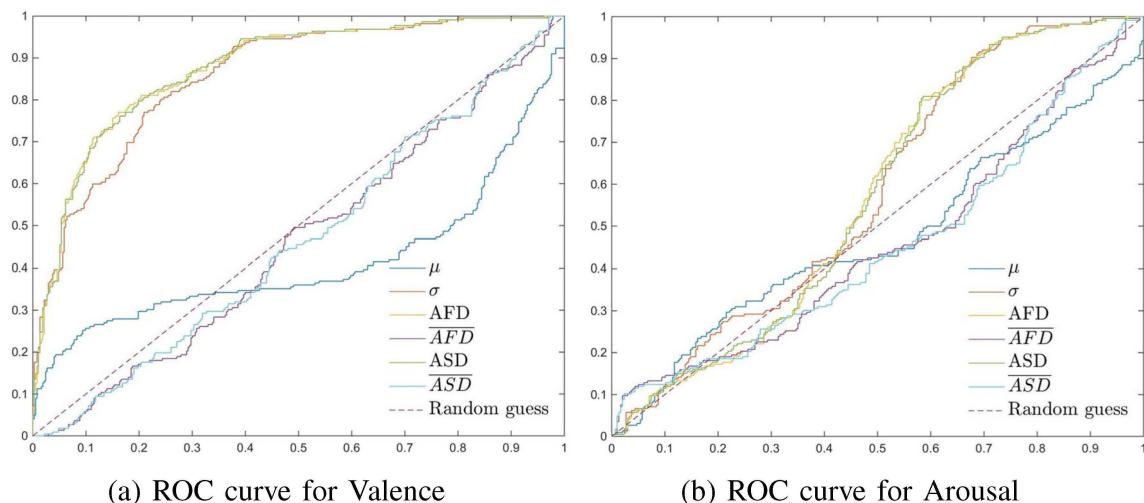


Fig. 4 ROC curves for each of the Statistical Features for Brainwaves

For both methods of feature extraction, the features associated with δ and θ performed better than the other band-waves. The best accuracy obtained was for the combination of the SPD for both δ and θ , resulting in 88.9% correctly classified instances.

The features that can be better associated with the affective state of Valence are \overline{ASD} , σ , \overline{AFD} and ASD , with 71.7, 70.7, 68.3 and 67.7%. Combining those features does not improve the results.

The highest classification accuracy rates were obtained using features extracted from the brainwaves, corroborating the neurophysiological theories that relate those with several different mental states. The Statistic features and Spectral Power Density represent the activity level in each bandwave and can give us an insight about the relation between the affective dimensions of Valence and Arousal and the brain activation in each frequency. In Figs. 4 and 5, we can see the Receiver Operating Characteristic (ROC) curve for both Statistic features and SPD respectively.

The red dashed line represents the equivalent of a random guess. The higher the curves are from this diagonal, the more sensitive it is regarding the class, Valence or Arousal. Analysing those curves for Valence we can see in Fig. 4a that σ , AFD and ASD have the best results, as well as δ and θ in Fig. 5a, corroborating the results obtained from both classification methods. The results for Arousal show curves close to the diagonal, again corroborating our previous results, of low accuracy for all methods in general.

Figures 6 and 7 show the distribution of features providing worst accuracy, in Fig. 6 and best accuracy, in Fig. 7. We can see that the features of the methods with worst accuracy, such as the HOC features and the Statistical features in time domain, overlap for the classes of High and Low Valence and Arousal. On the contrary, there is less overlap on the features obtained from the methods with best accuracy, such as σ and SPD of Brainwaves.

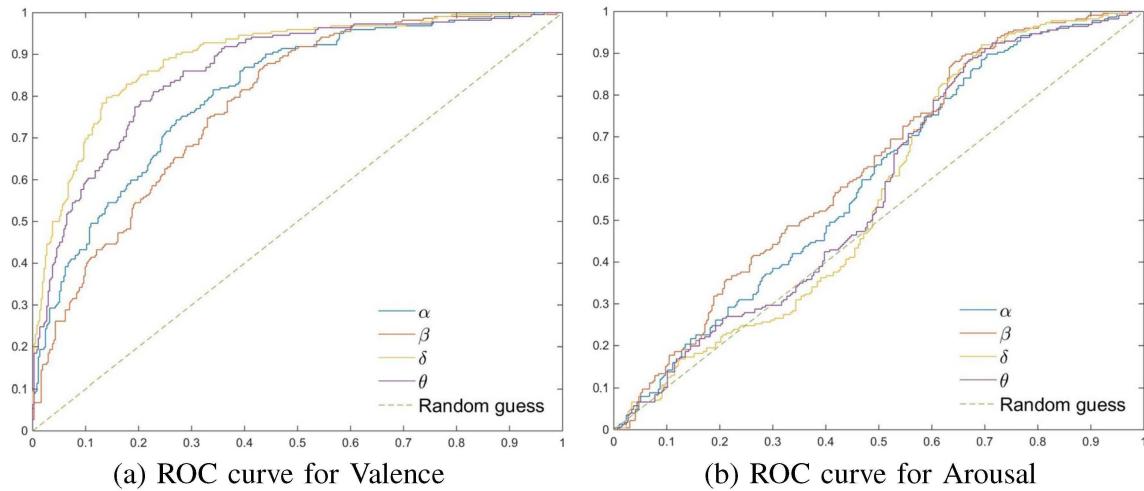


Fig. 5 ROC curves for the SPD of each of the Brainwaves.

Even though we obtained good results for some features, we can see in Figs. 6 and 7 that histograms of even the best features overlap considerably and result in ROC curves close to Random guess, as seen in Figs. 4 and 5. This characteristics observed amongst the features investigated could be due to many reasons. The sensitivity of the self-assessment scale used to garner affect ratings is subjective, as it is based on the thoughts and impressions of the participant about the video he/she watched. Moreover, it is often the case that people do not know how to articulate their actual emotions and associated states due to ambiguity and mixed mental activities [39].

Therefore, it is potentially the case that some of the participants could not precisely entail their actual emotional state using the SAM scale. Due to this factor, classification models were generated twice using two different mapping

schemes in order to determine the impact from ambiguous annotations that potentially arise from the selection of Valence and Arousal values from the middle of the self-assessment scale. As the results indicated, placing such a constraint on the ranges of affect to be modelled improved the overall classification performance.

In the majority of the investigations, the classification accuracies obtained for Valence outperformed those obtained for Arousal. It is difficult to determine why this was the case but several factors may have contributed to this effect. One possible reason is that the concept of Arousal may be more difficult to understand and categorize than Valence, resulting in inconsistent labelling. In addition, participants within the DEAP dataset watched video clips as a stimuli, hence were passive during that time, resulting in a small range of Arousal values that were not distinctive

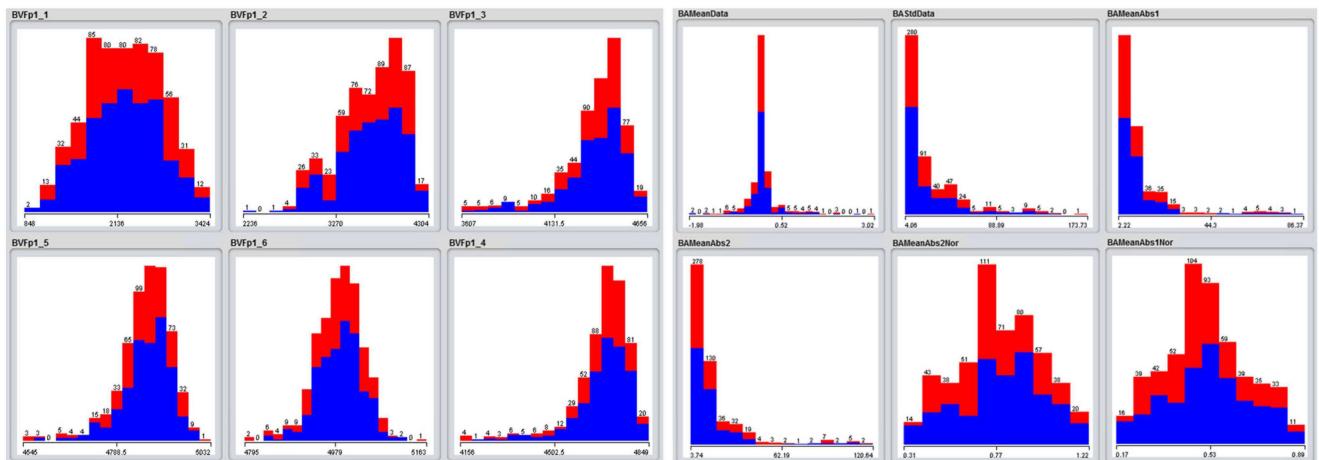


Fig. 6 Histograms of the features providing worst accuracy using Bipartition: the High class in red and Low in blue

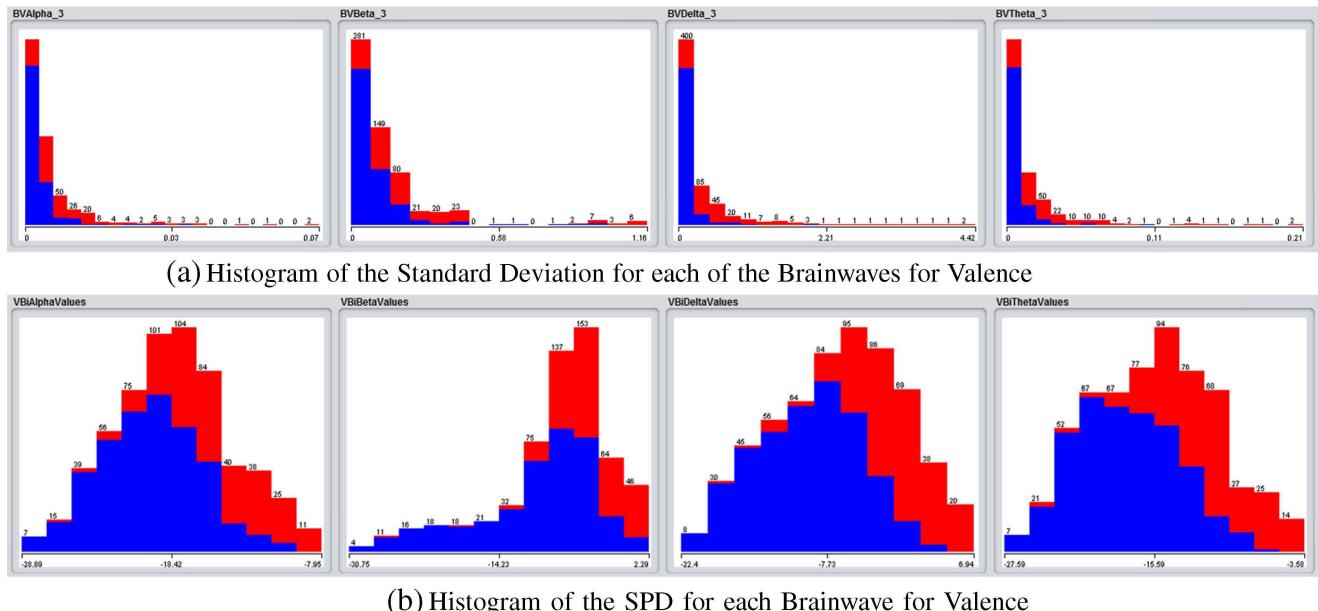


Fig. 7 Histograms of the features providing best accuracy using Bipartition: the High class in red and Low in blue

enough to be picked up by the classifier. This specific aspect could be improved if the data were obtained using a virtual environment, where the person has a greater sense of presence, hence having more influence in their emotional state, as discussed in chapter 2.

6 Conclusions and future work

This paper investigated exploiting electroencephalogram data as an input modality for the purpose of providing VEs with the ability to recognize and detect the emotional states of users. Consequently, the results from several experiments using different sets of features, especially the ones descendant from brainwaves, extracted from EEG data within the DEAP dataset show the potential of utilizing EEG signal data.

In addition, the observed discrepancy in classification accuracy due to different affective state mapping schemes was discussed, indicating that a degree of ambiguity will exist within such datasets, which has an obvious effect on the ability to accurately model affective states.

Moreover, combining several features together does not necessarily increase classification accuracy, as discussed in chapter 4. For example, as shown in Table 4, Combining δ 's and θ 's statistical features for Arousal, gives the accuracy of 73.8%, worst than the result for δ only.

Additionally, as the results depict, the features extracted from α , β , δ and θ waves and the classification accuracies obtained for Valence makes it potentially suitable as a metric for measuring this aspect of the affective state of a

user, ranging from negative to positive (i.e. *Low-Valence* to *High-Valence*).

The preliminary results shown in this article will help informing and leading to further experiments that eventually integrate different input modalities together with EEG in order to potentially provide a more robust model of the user's affective state. The current set of investigations would benefit if repeated using another mapping scheme based on Fuzzy Logic, for example, in an effort to improve the classification of potentially ambiguous affective states.

It is also interesting to extend the investigation regarding brain activation and the affective dimensions of Valence and Arousal. Not only how the negative (*Low-Valence*) and positive (*High-Valence*) states relate in terms of absolute values with the brain activation, but also how this activation is propagated through the entire extension of the brain.

Nonetheless, it is important to expand the study and the methods to real-time applications, and determine how those might behave in the real scenario of VEs. Not only taking into account the computational cost, aiming for real-time and embedded systems; but also how the virtual environment should adapt to this new form of awareness and how the user will react to this new form of enhanced interaction.

The article also discusses the importance of taking into account the effective qualities of the virtual environment to improve user experience and the many potential applications of such awareness for a different range of areas, such as medicine, education, entertainment and life style. The affective qualities of a virtual environment contribute to the engagement or feeling of presence of the user and

vice-versa. When the affective qualities of the VE do not match the expectations of the user or the affective level of the situation being lived at the environment, it may have a negative effect on the user experience. Recognizing the importance of the affective qualities and awareness of a VE and introducing these often neglected aspects into the development process will improve the user experience.

Acknowledgements The authors would like to thank COST for supporting the work presented in this paper (COST-STSM-TD1405-33385) and CNPq for the Science Without Borders Scholarship.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Allison BZ (2010) Toward ubiquitous BCIs. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 357–387
- Aymerich-Franch L (2010) Presence and emotions in playing a group game in a virtual environment: the influence of body participation. *Cyberpsychol Behav Soc Netw* 13(6):649–54
- Bekele E, Wade J, Bian D, Fan J, Swanson A, Warren Z, Sarkar N (2016) Multimodal adaptive social interaction in virtual environment (masi-vr) for children with autism spectrum disorders (asd). In: Virtual Reality (VR), vol 2016. IEEE, pp 121–130
- Botella C, Quero S, Banos R, Perpina C, Garcia Palacios A, Riva G (2004) Virtual reality and psychotherapy. *Stud Health Technol Inform* 99:37–54
- Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. *J Behav Therapy Exp Psychiatr* 25(1):49–59
- Brahnam S, Jain LC (2010) Virtual reality in psychotherapy, rehabilitation, and disease assessment. Springer
- Bransford JD, Brown AL, Cocking RR (1999) How people learn: brain, mind, experience and school. National Academy Press, Washington, DC
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Chang CC, Lin CJ (2001) LIBSVM - A Library for Support Vector Machines
- Cotrina-Atencio A, Ferreira A, Filho TFB, Menezes MLR, Pereira CE (2012) Avaliação de técnicas de extração de características baseadas em power spectral density, high order crossing e características estatísticas no reconhecimento de estados emocionais. In: XXIII Congresso Brasileiro em Engenharia Biomédica – XXIII, CBEB, Porto de Galinhas, PE, Brazil
- Davidson RJ (1992) Anterior cerebral asymmetry and the nature of emotion. *Brain Cogn* 20(1):125–151
- Davidson RJ, Jackson DC, Kalin NH (2000) Emotion, plasticity, context, and regulation: Perspectives from affective neuroscience. *Psychol Bullet* 126(6):890–909
- Dermer A (2016) Relaxing at the perfect beach : influence of auditory stimulation on positive and negative affect in a virtual environment
- EL-Manzalawy Y (2005) WLSVM
- Fairclough SH, Gilleade K, Ewing KC, Roberts J (2013) Capturing user engagement via psychophysiology: Measures and mechanisms for biocybernetic adaptation. *Int J Auton Adapt Commun Syst* 6(1):63–79
- Glantz K, Rizzo A (2003) Virtual reality for psychotherapy: Current reality and future possibilities. *Psychotherapy*
- Hodges L, Anderson P, Burdea G, Hoffman H, Rothbaum B (2001) Treating psychological and physical disorders with vr. *IEEE Computer Graphics and Applications*
- Hu WL, Akash K, Jain N, Reid T (2016) Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine* 49(32):48–53
- Ip HHSI, Byrne J, Cheng SH, Kwok RCW (2011) The samal model for affective learning: A multidimensional model incorporating the body, mind and emotion in learning. In: DMS, Knowledge Systems Institute, pp 216–221
- Izquierdo-Reyes J, Ramirez-Mendoza RA, Bustamante-Bello MR, Navarro-Tuch S, Avila-Vazquez R (2017) Advanced driver monitoring for assistance system (admas). *International Journal on Interactive Design and Manufacturing (IJIDeM)* 11:1–11
- Jenke R, Peer A, Buss M (2014) Feature extraction and selection for emotion recognition from eeg. *IEEE Trans Affect Comput* 5(3):327–339
- Kedem B (1986). Spectral Analysis and Discrimination by Zero-Crossings. *Proceedings of the IEEE* 74
- Kim MK, Kim M, Oh E, Kim SP (2013) A review on the computational methods for emotional state estimation from the human eeg. *Computational and mathematical methods in medicine*
- Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I (2012) Deap: A database for emotion analysis and using physiological signals. *IEEE Trans Affect Comput* 3(1):18–31
- Konstantinidis EI, Frantzidis CA, Pappas C, Bamidis PD (2012) Real time emotion aware applications: A case study employing emotion evocative pictures and neuro-physiological sensing enhanced by graphic processor units. *Comput Methods Program Biomed* 107(1):16–27. Advances in Biomedical Engineering and Computing: the MEDICON conference case
- Lab L (2017). Second life. <http://secondlife.com/>
- Lee EAL (2011) An investigation into the effectiveness of virtual reality-based learning. PhD thesis, Murdoch University
- Lin CT, Lin FC, Chen SA, Shao-Wei T-CL, Ko CLW (2010a) Eeg-based brain-computer interface for smart living environmental auto-adjustment. *J Med Biol Eng* 30(4):237–245
- Lin YP, Wang CH, Jung TP, Wu TL, Jeng SK, Duann JR, Chen JH (2010b) Eeg-based emotion recognition in music listening. *IEEE Trans Biomed Eng* 57(7):1798–1806
- Murugappan M, Ramachandran N, Sazali Y et al (2010) Classification of human emotion from eeg using discrete wavelet transform. *J Biomed Sci Eng* 3(04):390
- Niedermeyer E, Da Silva FL (1993) *Electroencephalography: Basic principles, clinical applications, and related fields*. Williams & Wilkins, Baltimore, MD
- Niemiec CP (2002) Studies of emotion: A theoretical and empirical review of psychophysiological studies of emotion. In: *Journal of Undergraduate Research*, University of Rochester, vol 1, pp 15–18
- Nijholt A, Tan D, Pfurtscheller G, Brunner C, Millán JdR, Allison B, Graimann B, Popescu F, Blankertz B, Müller KR (2008) Brain-computer interfacing for intelligent systems. *IEEE Intell Syst* 23(3):72–79

34. Olszewski K, Lim JJ, Saito S, Li H (2016) High-fidelity facial and speech animation for vr hmds. *ACM Trans Graph (TOG)* 35(6):221
35. Ontiveros-Hernández NJ, Pérez-Ramírez M, Hernández Y (2013) Virtual reality and affective computing for improving learning. *Res Comput Sci* 65:121–131
36. Parsons TD, Rizzo A (2008) Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias A meta-analysis. *J Behav Therapy Exp Psychiatry* 39:250–261
37. Petrantonakis PC, Hadjileontiadis LJ (2010a) Emotion recognition from eeg using higher order crossings. *IEEE Trans Inf Technol Biomed* 14(2):186–197
38. Petrantonakis PC, Hadjileontiadis LJ (2010a) Emotion recognition from eeg using higher order crossings. *IEEE Trans Inf Technol Biomed* 14(2):186–197
39. Picard RW (2003) Affective computing: Challenges. *Int J Human-Comput Stud* 59(1-2):55–64
40. Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans Pattern Anal Mach Intell* 23(10):1175–1191
41. Picard RW, Papert S, Bender W, Blumberg B, Breazeal C, Cavallo D, Machover T, Resnick M, Roy D, Strohecker C (2004) Affective learning — a manifesto. *BT Technol J* 22(4):253–269
42. Riva G, Mantovani F, Capideville C, Preziosa A, Morganti F, Villani D, Gaggioli A, Botella C, Alcañiz M (2007) Affective interactions using virtual reality: the link between presence and emotions. *Cyberpsychology Behav* 10(1):45–56
43. Ruscher G, Kruger F, Bader S, Kirste T (2011) Controlling smart environments using brain computer interface. Proceedings of the 2nd Workshop on Semantic Models for Adaptive Interactive Systems
44. Russell JA (1980) A circumplex model of affect. *J Personal Soc Psychol* 39(6):1161–1178
45. Schlägl A, Slater M, Pfurtscheller G (2002) Presence research and eeg. In: Proceedings of the 5th International Workshop on Presence, vol 1, pp 9–11
46. Steinwart I, Christmann A (2008) Support vector machines. Springer Science and Business Media
47. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J Chem Inf comput Sci* 43(6):1947–1958. doi:[10.1021/ci034160g](https://doi.org/10.1021/ci034160g)
48. Szafir D, Mutlu B (2012) Pay attention!: Designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, NY, USA, CHI '12, pp 11–20
49. Wolpaw JR (2012) Brain-computer interfaces: Progress, problems, and possibilities. In: Proceedings of the 2Nd ACM SIGCHI International Health Informatics Symposium. ACM, NY, USA, pp 3–4

Design and Implementation of an Emotion Analysis System Based on EEG Signals

ZHANG Yutian, HUANG Shan, ZHANG Jianing, FAN Ci'en

School of Electronic Information, Wuhan University, Wuhan Hubei 430072, China

Abstract: Traditional brain-computer systems are complex and expensive, and emotion classification algorithms lack representations of the intrinsic relationships between different channels of electroencephalogram (EEG) signals. There is still room for improvement in accuracy. To lower the research barrier for EEG and harness the rich information embedded in multi-channel EEG, we propose and implement a simple and user-friendly brain-computer system for classifying four emotions: happiness, sorrow, sadness, and tranquility. This system utilizes the fusion of convolutional attention mechanisms and fully pre-activated residual blocks, termed Attention-Convolution-based Pre-Activated Residual Network (ACPA-ResNet). In the hardware acquisition and preprocessing phase, we employ the ADS1299 integrated chip as the analog front-end and utilize the ESP32 microcontroller for initial EEG signal processing. Data is wirelessly transmitted to a PC through UDP protocol for further preprocessing. In the emotion analysis phase, ACPA-ResNet is designed to automatically extract and learn features from EEG signals, thereby enabling accurate classification of emotional states by learning time-frequency domain characteristics. ACPA-ResNet introduces an attention mechanism on the foundation of residual networks, adaptively assigning different weights to each channel. This allows it to focus on more meaningful EEG signals in both spatial and channel dimensions while avoiding the problems of gradient dispersion and explosion associated with deep network architectures. Through testing on 16 subjects, our system demonstrates stable EEG signal acquisition and transmission. The novel network significantly enhances emotion recognition accuracy, achieving an average emotion classification accuracy of 95.1%.

Key words: electroencephalogram(EEG); deep learning; emotional analysis; ACPA ResNet; attention mechanism

Electroencephalogram (EEG) signals are weak electrical signals generated by neuronal activities in the brain, reflecting neural activities and playing a crucial role in fields such as medicine, neuroscience, and psychology^{[1][2]}. Recently, EEG signals have attracted significant attention in the realm of emotion recognition research, as they not only reflect various brain activities but also effectively indicate human emotional states. However, EEG signals are low in amplitude, generally between 0.02-0.5μV when collected via non-invasive brain-computer interfaces, and their frequency and amplitude can change in response to external stimuli, thought processes, or changes in physical states^[3]. Therefore, EEG acquisition devices must be highly precise. Additionally, to facilitate the collection of EEG signals, the devices should also be compact and portable.

Presently, deep neural networks (DNN) are widely applied for feature extraction and have achieved commendable results in processing images, videos, voices, and texts, outperforming traditional algorithms^{[4][5]}. However, in EEG research, traditional emotion analysis methods still rely on manually designed feature extractors. For instance, Kumar et al. ^[6] used Linear Kernel Least Squares Support Vector Machines (LS-SVM) and Back Propagation Artificial Neural Networks (BP-ANN) for binary classification of emotions based on valence and arousal models, achieving accuracies of 61.17% and 64.84%, respectively. Chen et al. ^[7] proposed a method combining Data Space Adaptation (DSA) and Common Spatial Patterns (CSP) for EEG-based emotion classification, reaching an accuracy of 68.3% in a two-day emotion classification involving 12 subjects over five days. Compared to these, DNNs possess the ability to automatically extract and learn data features, making emotion analysis more accurate and efficient. For example, Li Youjun et al. ^[8] applied Stacked Denoising Autoencoder (SAE) and an LSTM-based Recurrent Neural Network (RNN) to recognize emotions from mixed physiological signals including EEG, achieving an accuracy of 79.26%. Li C et al. ^[9] introduced a hybrid neural network combining CNN, DNN, and LSTM, which achieved an accuracy of 75.52% on the EEG Movement/Imagery dataset. Zhou Yijun et al. ^[10] fused MFCC features extracted from wavelet-transformed EEG signals and utilized the features of a deep residual network (ResNet18) for emotion classification, achieving arousal and valence recognition rates of 86.01% and 85.46% in the DEAP database. Nonetheless, applying DNNs to EEG-based emotion analysis systems still poses certain challenges, and there is room for improvement in recognition accuracy, necessitating further research and optimization^[11].

In summary, current EEG signal collection and emotion analysis techniques still face deficiencies in data transmission, device portability, and accuracy of emotion analysis. Therefore, researching a real-time, portable, and highly accurate EEG signal collection and emotion analysis system is particularly crucial. Through efficient and accurate EEG signal collection and emotion classification, researchers can deepen their understanding of the relationship between emotions and EEG signals, advancing the field of emotion analysis and providing new avenues for the treatment and intervention of neurological disorders.

1 System Design

To address the shortcomings in EEG signal collection related to data transmission, device portability, and research costs, as well as deficiencies in emotion classification analysis, this paper aims to design and implement a simple, portable, affordable, and highly accurate emotion analysis system based on EEG signals.

The system is divided into two main parts: the front-end hardware collection and preprocessing section (Figure 1), and the back-end emotion classification algorithm section (Figure 4). Subjects generate different emotions while watching videos. The signals are collected and amplified by the analog front-end, then wirelessly transmitted via an MCU to the PC for further preprocessing and emotion classification.

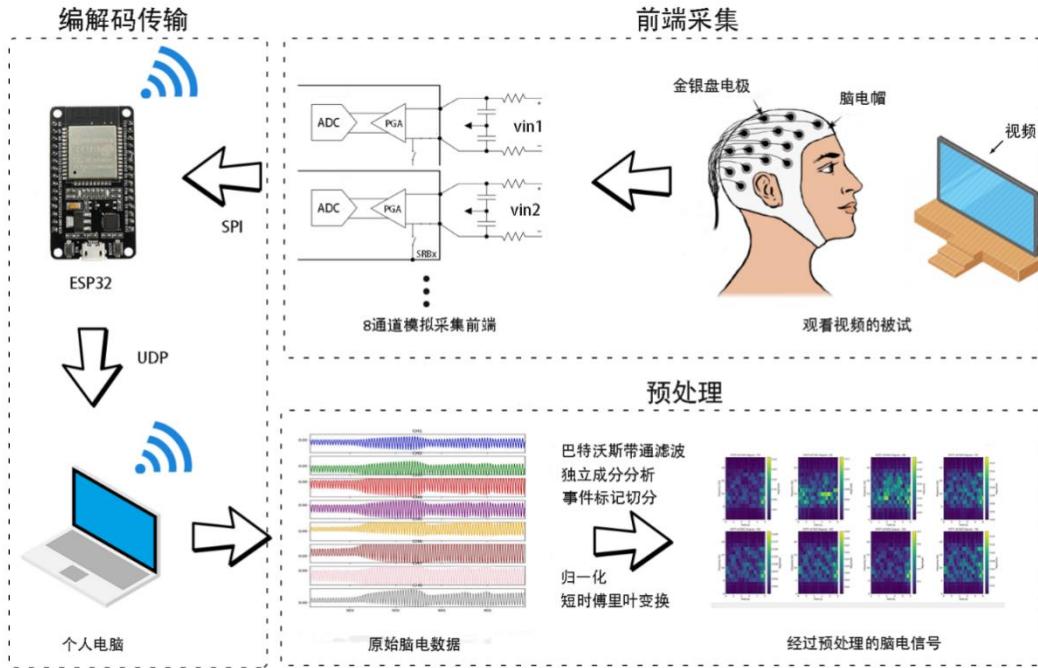


Fig.1 Frontend hardware acquisition and preprocessing process

1.1 Frontend Hardware Collection and Preprocessing

1.1.1 Frontend Collection

The collection of EEG signals relies on the accurate placement of electrodes. This system utilizes the international 10-20 system (Figure 3) as the reference standard, selecting eight electrode channels: P3, Pz, P4, O1, Oz, O2, T5, and T6.

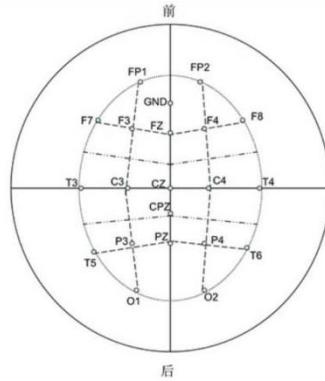


Fig.2 International 10-20 System

These electrode channels cover various areas on the posterior scalp, allowing the collected EEG signals to comprehensively reflect the subject's emotional state. Additionally, a ground electrode and a bias electrode are introduced, connected to GND and the earlobes respectively, to enhance the common mode rejection ratio and reduce interference caused by subject movement. The electrodes chosen are gold disc electrodes, and a suitable amount of conductive gel is applied between the electrodes and the subject's scalp to minimize environmental disturbances, enhance system stability, and ensure high-quality EEG signal collection.

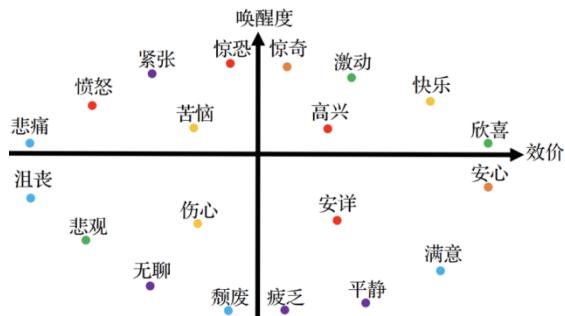


Fig.3 Two-dimensional emotion classification model

After the subjects don the collection device, they generate four types of emotions—happiness, distress, sadness, and calmness—by watching related videos and through self-reflection. The choice of emotions is guided by typical values determined by the valence-arousal (VA) model^[12], a two-dimensional continuous model comprised of arousal and valence levels, as illustrated in Figure 3.

Given that the EEG signals collected via non-invasive brain-computer interfaces typically range from 0.02 to 0.5 μV and are prone to disturbances, the system employs an eight-channel, 24-bit high-precision AD chip—ADS1299—as the core component of the analog frontend. After sampling, the digital signals are transmitted to the MCU via the SPI communication protocol.

1.1.2 Data Transmission

To enhance the system's portability, enable wireless transmission of EEG signals, and reduce research costs, the system uses a domestically produced MCU equipped with a 2.4GHz WIFI chip—ESP32—as the core component for transmitting and receiving EEG data. The MCU communicates with the analog frontend via SPI to receive eight channels of 24-bit data. These data are then decoded into eight floating-point numbers in physical units (μV), re-encoded, and transmitted in real-time to the PC listening via the UDP protocol, which is minimal in overhead and reliable. Experiments show that the data transmission delay is as low as 0.02 ms, with no packet loss observed.

1.1.3 Preprocessing and Feature Extraction

The raw EEG signals received by the PC are mixed with multiple frequencies and include power frequency interference and artifacts such as eye movements, making them unsuitable for direct use. The system undertakes several preprocessing and feature extraction steps. It employs an 11th-order Butterworth band-pass digital filter to extract the 5-18 Hz frequency band (α band) from the raw EEG signals. Event marking and segmentation are followed by Independent Component Analysis (ICA) to remove artifacts related to eye movements. After normalization, the short-time Fourier transform (STFT) is used to extract time-frequency domain features from the artifact-free 5-18 Hz EEG signals. The dimensionally enhanced signals are then fed into a neural network for emotion classification.

1.2 Emotion Classification Algorithm

After the preprocessing phase, the PC receives the feature data of EEG signals, which then enters the emotion classification phase. The system incorporates a neural network that combines convolutional attention mechanisms and fully pre-activated residual blocks, known as the Attention-Convolution-based Pre-Activated Residual Network (ACPA-ResNet), to perform emotion classification. The specific architecture of this network is shown in Figure 4.

This system architecture supports efficient processing and classification of emotional states from EEG signals, utilizing advanced signal processing techniques and neural network models to ensure high accuracy and reliability in emotion detection.

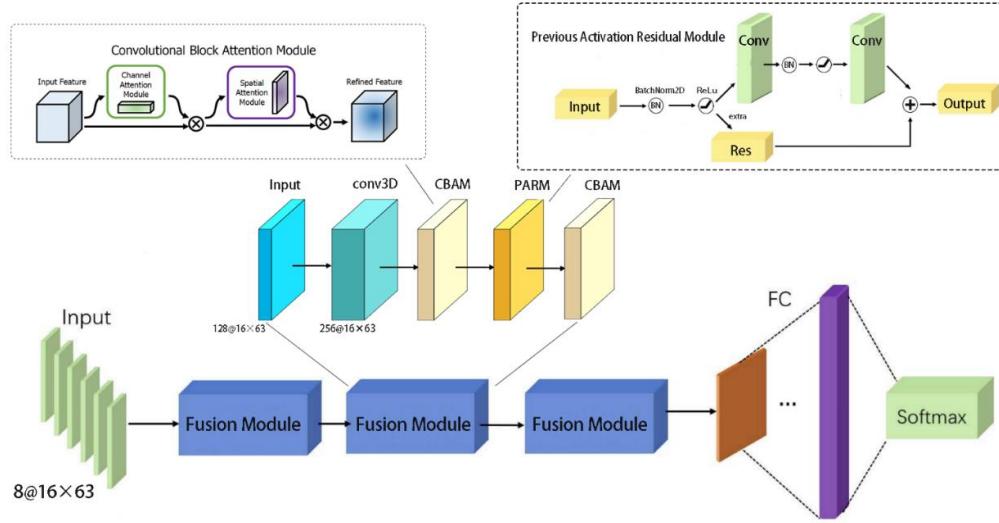


Fig.4 ACPA ResNet Network Architecture

Input Layer: This layer receives input in the form of two-dimensional features from eight channels, each channel providing a 16×63 matrix. These inputs are passed to the Fusion Module.

Hidden Layers — Fusion Module: This module consists of three parts: a convolutional layer, a Convolutional Block Attention Module (CBAM)^[13], and a Fully Pre-Activated Residual Module (PARM). The convolutional layer increases the number of channels to facilitate the subsequent channel-wise attention mechanism in the CBAM. The CBAM is a simple yet effective attention module that computes attention weights in both spatial and channel dimensions, and then adapts the feature maps by multiplying them with the original feature maps, aligning well with the characteristics of EEG signals. The PARM is an improvement over the traditional Residual Block (ResBlock)^[14]. Unlike the original ResBlock pattern of weight-BatchNorm-ReLU-weight-BatchNorm^[15], the Pre-activation ResBlock uses BatchNorm-ReLU-weight-BatchNorm-ReLU-weight. This configuration allows gradients to flow unimpeded through fast connections to any layer in the module, thus accelerating convergence and ensuring robust performance.

Hidden Layer — Fully Connected Layer (FC): This layer uses the ReLU activation function to introduce non-linearity, enhancing the ability to capture complex patterns in the data.

Output Layer: Comprises a fully connected layer with four neurons, each outputting the prediction results for one of the four emotional states analyzed (happiness, distress, sadness, and calmness). This layer is responsible for mapping the high-level features learned by the network into the final emotion classifications.

2 Experiments and Results Analysis

2.1 Experimental Setup

The hardware setup for this experiment, as shown in Figure 5, includes an EEG cap, eight-channel gold disc electrodes, the ADS1299 acquisition module, an SPI communication adapter board, and the ESP32 module. A deep learning framework based on Python 3.7 and PyTorch 1.7.1 was established, utilizing a Lenovo Legion Y7000P with an NVIDIA GeForce RTX 2060 for training. The model was trained using ten-fold cross-validation, with 855 samples for training and the remaining 95 for testing. The average of the ten folds was taken as the recognition result for each subject. Sixteen volunteers were recruited as subjects, including individuals of different genders and emotional states. The model's recognition performance was assessed using the average recognition rate and standard deviation among these 16 subjects.



Fig.5 Physical diagram of front-end hardware acquisition system

2.2 EEG Signal Acquisition Results

The ADS1299 analog front-end successfully captured high-quality EEG signals from subjects. Figure 6 displays the raw EEG signals from a subject, as well as the signals after preprocessing and application of the Short Time Fourier Transform (STFT), showing clearly distinguishable EEG waveforms.

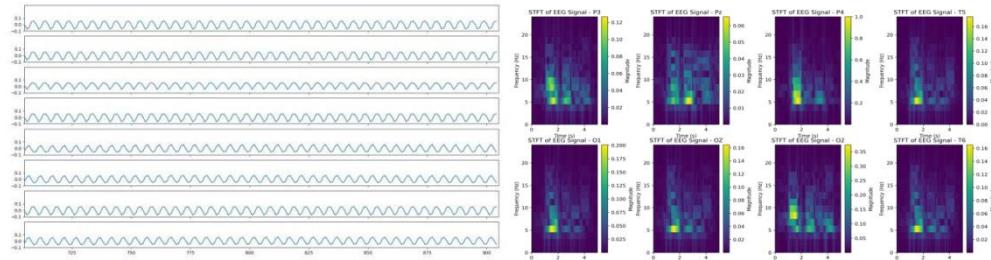


Fig.6 Original EEG signals and EEG signals passing through STFT

2.3 Accuracy of Emotion Analysis Algorithm

The emotion classification model, constructed using ACPA-ResNet, classified emotional states from the collected EEG signals and was compared with existing studies. System recognition accuracy was evaluated against an emotion status survey filled out by subjects during the experiments. The results, as shown in Figure 7 and Table 1, demonstrate that the system performs excellently in terms of emotion analysis accuracy, successfully classifying different emotional states.

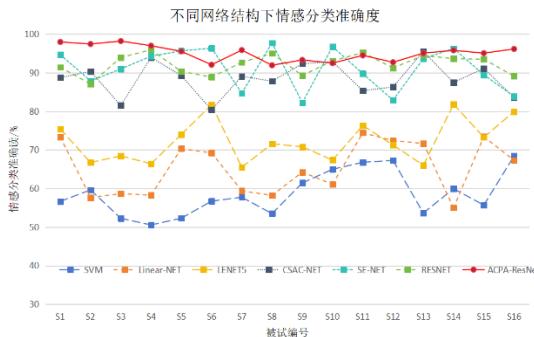


Fig.7 Accuracy of sentiment classification under different network structures

Table 1 Comparison of the accuracy of emotion recognition between this method and other methods

方法	准确度
SVM	58.6%
Linear-NET	65.3%
LENET5	72.3%
CSAC-NET	88.5%
SE-NET	91.1%
RESNET	92.2%
ACPA-RESNET	95.1%

3 Conclusion

This study designed and implemented an emotion analysis system based on EEG signals using the ADS1299 analog front-end and the ESP32 microcontroller. Data transmission through the UDP protocol ensured real-time and stable performance. The experiments confirmed the system's ability to obtain high-quality EEG signals and to accurately reflect the brain's activity under different emotional states. The system demonstrated excellent real-time performance and stable data transmission capabilities. It showed outstanding accuracy in emotion analysis, accurately assessing subjects' emotional states. However, there are still limitations to the system, and future improvements could focus on optimizing algorithms and hardware design, expanding the range of test subjects, and diversifying emotion state classifications. This research makes a positive contribution to the field of EEG signal collection and emotion analysis, offering new ideas and direction.

- [1] WANG H Y ,HU J F,WANG Y L. A review of EEG signal processing methods[J]. Computer Era,2018(1):13-15,19.
- [2] JIANG J F,ZENG Y,LIN Z M,et al. Review on EEG-Based Emotion Assessment[J]. Journal of Information Engineering University,2016,17(6):686-693.
- [3] ZHANG H J,WANG H C. Research on classification and recognition of multi lead EEG signals [J]. Computer Engineering and Applications, 2008,44 (24): 228-230
- [4] LI J Y,DU X B,ZHU Z J,et al. Deep Learning for EEG-based Emotion Recognition: A Survey.Journal of Software, 2023, 34(1): 255-276.
- [5] ZHANG C,GUO Y,LI M. Review of Development and Application of Artificial Neural Network Models[J]. Computer Engineering and Applications,2021,57(11):57-69.
- [6] Kumar N,Khaund K,Hazarika S M. Bispectral analysis of EEG for emotion recognition[J]. Procedia Computer Science,2016,84:31-35.
- [7] CHEN J X,ZHENG R,ZHANG P W,et al. Electroencephalogram Emotion Classification Based on Data Space Adaptation and Common Spatial Pattern. Computer Engineering,2019,45(4):296-301.
- [8] LI Y J,HUANG J J,WANG H Y,et al. Study of emotion recognition based on fusion multi-modal bio-signal with SAE and LSTM recurrent neural network[J]. Journal on Communications,2017,38(12):109-120.
- [9] Li C, Yang H H, Wu X, et al. Improving EEG-Based Motor Imagery Classification Using Hybrid Neural Network[C]. IEEE 9th International Conference on Information, Communication and Networks (ICICN), 2021: 486-489
- [10] ZHOU Y J, LI D D,WANG Z, et al. Cepstrum feature fusion for EEG emotion classification. Computer Engineering and Applications,2020,56(21):164-169.
- [11] ZHOU Y J, LI D D,WANG Z, et al. Cepstrum feature fusion for EEG emotion classification. Computer Engineering and Applications,2020,56(21):164-169.
- [12] QING T P, SHENG H, YUE L,et al. Survey of Research on EEG Signal Emotion Recognition[J]. Computer Engineering and Applications,2023,59(15):38-54.
- [13] Woo, Sanghyun, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. “Cbam: Convolutional Block Attention Module.” In Proceedings of the European Conference on Computer Vision (ECCV), 3 – 19.
- [14] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770-778, 2016
- [15] GUO Y X,YANG W,LIU Q,et al. Survey of residual network[J]. Application Research of Computers,2020,37(5):1292-1297.

CLDTA: Contrastive Learning based on Diagonal Transformer Autoencoder for Cross-Dataset EEG Emotion Recognition

Yuan Liao, Yuhong Zhang, Shenghuan Wang, Xiruo Zhang, Yiling Zhang, Wei Chen, Yuzhe Gu, and Liya Huang*

Abstract—Recent advances in non-invasive EEG technology have broadened its application in emotion recognition, yielding a multitude of related datasets. Yet, deep learning models struggle to generalize across these datasets due to variations in acquisition equipment and emotional stimulus materials. To address the pressing need for a universal model that fluidly accommodates diverse EEG dataset formats and bridges the gap between laboratory and real-world data, we introduce a novel deep learning framework: the Contrastive Learning based Diagonal Transformer Autoencoder (CLDTA), tailored for EEG-based emotion recognition. The CLDTA employs a diagonal masking strategy within its encoder to extracts full-channel EEG data's brain network knowledge, facilitating transferability to the datasets with fewer channels. And an information separation mechanism improves model interpretability by enabling straightforward visualization of brain networks. The CLDTA framework employs contrastive learning to distill subject-independent emotional representations and uses a calibration prediction process to enable rapid adaptation of the model to new subjects with minimal samples, achieving accurate emotion recognition. Our analysis across the SEED, SEED-IV, SEED-V, and DEAP datasets highlights CLDTA's consistent performance and proficiency in detecting both task-specific and general features of EEG signals related to emotions, underscoring its potential to revolutionize emotion recognition research.

Index Terms—EEG, Emotion Recognition, Contrastive Learning, Transfer Learning, Cross-datasets

I. INTRODUCTION

Recently, the domains of human-computer interaction and affective computing have seen substantial advancements due to the exploration of emotion recognition [1]. Compared to facial expressions, movements, and linguistic cues, electroencephalogram (EEG) provides a

This work was supported by the National Natural Science Foundation of China (Grant No. 61977039), “New Infrastructure Development & University Informatization” research project of China Association for Educational Technology (Grant No. XJJ1202205007) and Open subject of cognitive EEG and transcranial, electrical stimulation regulation of neuraile (Grant No. BRKOT-NJUPT-20220630H). (*Corresponding author: Liya Huang*).

Yuan Liao, Xiruo Zhang, Yiling Zhang, Wei Chen, Yuzhe Gu and Liya Huang are with college of electronic and optical engineering & college of flexible electronics (future technology), Nanjing University of Posts and Telecommunications, Jiangsu, 210023, China, E-mail: {1022020619; 1022020620; 1022020621; 1222025223; 1222025429; huangly}@njupt.edu.cn.

Yuhong Zhang is with the Department of Bioengineering, University of California, San Diego, La Jolla, 92093, USA, email: yuz291@ucsd.edu.

Shenghuan Wang is with the College of Letters and Science, UC Davis, Davis, CA 95618, USA (e-mail: dvswang@ucdavis.edu).

more direct and objective measurement of human emotional responses [2], [3]. It boasts high temporal resolution and is difficult to fake or conceal [4], [5]. Moreover, in contrast to other neural imaging modalities such as fMRI and ECOG, EEG is non-invasive and relatively easy to collect, which has led to an increasing focus on EEG-based emotion recognition in the field of Brain-Computer Interface [6], [7], [8]. This interest has spurred the development and launch of affordable, gel-free wireless EEG devices, alongside studies employing a few electrodes for detecting emotions and depression [9], [10], [11]. In these studies, deep learning methods have achieved significant results in EEG-based emotion recognition. However, a pressing issue is that current deep learning models, despite performing well on various datasets, need different parameters set for each (e.g., SEED [12] and DEAP [13]) to accommodate distinct EEG data configurations. This laborious and time-consuming retraining process significantly hampers progress in emotion recognition using EEG in real-world scenarios.

Extensive research has been conducted on the EEG representation of emotions [14], [15]. Current emotion recognition methodologies widely apply Differential Entropy (DE) features [16]. It has also been found that analyzing functional connectivity is essential for the advancement of emotion recognition. Based on this, graph neural networks (GNNs) [17] and convolutional neural networks (CNNs) [18] have been proposed to extract spatial embedding of DE features among different EEG channels. Furthermore, long short-term memory (LSTM) [19] and attention mechanisms have been utilized to learn emotion-related EEG representations [20]. These approaches leverage the end-to-end capabilities of deep neural networks, thereby eliminating the need for manual feature extraction.

Though effective, these methodologies heavily rely on two key assumptions: data quality and quantity. Firstly, the quality of data from consumer-grade EEG devices often falls short compared to that from strictly controlled laboratory environments, resulting in few models performing well across both dataset types [21]. Secondly, popular models such as GNNs and CNNs are typically designed for specific EEG datasets, which prevents current deep learning models from acting as a bridge to connect multiple datasets and facilitate the sharing of knowledge [17], [22]. Moreover, due to subject variability, current research primarily focuses on collecting extensive data for each participant and conducting lengthy

> TAFFC-2024-03-0186 <

training [3] to learn emotion patterns based on individual-specific EEG representations [9], [13], necessitating periodic model recalibration to ensure stable accuracy.

The scenario highlights the value of transfer learning (TL) for EEG-based emotion recognition, facilitating improvement through knowledge transfer. Key contributions include the BiDANN model by Li et al. [23], focusing on generalizability and feature identification; the RGNN by Zhong et al. [24], aimed at cross-subject variation and noise reduction; and the PR-PL framework by Zhou et al. [25], designed for more accurate, individualized recognition by minimizing label dependence. These TL approaches address core challenges in affective BCI, demonstrating significant advancements in emotion detection from EEG signals. The primary issue with existing approaches is their focus on applying transfer learning within a single dataset without checking if the learned features work across various datasets. Furthermore, these approaches typically use all data from the target domain[26], meaning all EEG records for new subjects must be available before transferring knowledge. This is impractical in real-world situations where data may be limited and quick emotion detection is necessary.

We introduce a novel contrastive pretraining transfer learning framework, named Contrastive Learning based on Diagonal Transformer Autoencoder (CLDTA), to enhance the performance of emotion recognition from EEG data in real-world scenarios. Our approach draws inspiration from BERT's Masked Language Modeling (MLM), where we simulate full-channel laboratory data and limited-channel real-world data by masking portions of EEG channels. Diagonal masking strategy and information separation technique trains the model to identify emotional representations that are independent of the number of EEG channels, thereby improving the model's applicability across various data settings. Furthermore, by comparing EEG signal samples from the same or different subjects, our model uncovers more generalizable emotional representations, independent of the subjects [22]. Knowledge transfer is then applied to utilize the learned emotional features and model parameters in real-world emotion recognition tasks.

The CLDTA framework is structured into two main stages: the contrastive learning process and the calibration-prediction process. Initially, the Diagonal Transformer Autoencoder (DTA) learns to represent emotions from EEG signals. Contrastive learning is then employed to amplify the alignment of features corresponding to identical emotions and to diminish the alignment of features corresponding to differing emotions. Subsequently, in the calibration-prediction stage, the pre-trained DTA, coupled with a newly initialized classifier, is fine-tuned using a small set of labeled samples from new subjects, ensuring swift personalization. Post-calibration, the model is equipped to perform precise emotion classification from EEG data. The integration of diverse data augmentation techniques during the contrastive learning phase significantly enhances the model's robustness and applicability across various datasets.

In summarize, the CLDTA model offers several distinct advantages:

- **Universality:** By implementing a diagonal masking strategy, the model can effectively learn brain network knowledge from high-quality, full-channel EEG datasets and apply it to realistic, relatively noisy, and lesser-channel EEG datasets. This enhances the model's universality, making it suitable for various data acquisition devices and aBCI use scenarios.
- **Rapid Adaptation:** By integrating contrasting learning and transfer learning mechanisms, the model rapidly adjusts to new subjects with minimal samples. This swift adaptation eliminates the need for extensive training, providing significant benefits for real-time emotion recognition applications.
- **Interpretability:** As a result of an information separation mechanism, the model converts EEG signals into understandable structures, facilitating visualization and analysis of individual emotion attributes. This innovation helps mitigate the common 'black-box' issue associated with deep learning models.
- **Validated Effectiveness:** Due to various data augmentation mechanisms, the robustness and accuracy of CLDTA have been validated across four publicly available EEG datasets: SEED, SEED-IV [27], SEED-V [28].

The structure of this paper is organized as follows: Section II outlines the related work, providing context and background for our study. Section III describes the methodology, including the development and implementation of our model. Section IV details the experimental setup, data collection, and evaluation metrics. Section V presents the results and offers an analysis of the findings. Finally, Section VI concludes the paper with a discussion of the implications, limitations, and future directions for research in this area.

II. RELATED WORKS

A. EEG-based Emotion Recognition

EEG-based emotion recognition involves feature extraction and classification, traditionally leveraging discrete wavelet transform (DWT), power spectral density (PSD), differential entropy (DE), and differential asymmetry (DASM) with SVM or LDA classifier [29].

Compared to conventional machine learning algorithms, deep learning has introduced end-to-end approaches that autonomously extract features using CNN, LSTM, GNN. For instance, Wang et al. [30] proposed a self-supervised EEG emotion recognition model based on CNN to enhance resource utilization efficiency. Ma et al. [31] developed a multimodal residual LSTM (MM-ResLSTM) network, while Song et al. [17] proposed a dynamic graph convolutional neural network (DGCNN) for EEG emotion recognition. More recently, the advent of Transformer models [32] achieved significant success in fields such as Natural Language Processing and

> TAFFC-2024-03-0186 <

Computer Vision. The emergence of Transformers also represents a significant evolution in discerning emotional states. Wang et al. [33], used attention mechanisms to focus on key features, helping to classify emotions by combining data from different parts of the brain.

B. Transfer Learning in EEG Processing

The high variability in individual EEG signals [34] limits the generalizability of deep learning methods in emotion recognition, confining many models to lab settings despite potential wider applications [35]. Transfer learning, aimed at applying knowledge from one domain to another, has shown promise in EEG analysis, especially in cross-session, cross-subject, and cross-database scenarios [36]. Research has primarily focused on cross-session and cross-subject scenarios to mitigate EEG signal variability over time and between individuals. Zhang et al. [37] introduced a similarity-guided transfer learning method using Maximum Mean Discrepancy (MMD) and TrAdaBoost for closer data distribution alignment. Domain adaptation (DA) techniques like the bi-hemispheres domain-adversarial neural network (Bi-DANN) [23] and regularized graph neural network (RGNN) aim to learn domain-invariant representations. Domain generalization (DG) methods, such as the two-phase prototypical contrastive domain generalization framework (PCDG) [38] and the Contrastive Learning method for Inter-Subject Alignment (CLISA) [22], reduce reliance on new subject data by identifying subject-invariant emotional representations. Li et al. [15] proposed a graph-based multi-task self-supervised learning model (GMSS) for more general representation learning.

In cross-database scenarios, addressing differences between databases remains challenging but crucial for model adaptability. Lin et al. [39] developed a personalized model using robust principal component analysis (RPCA) to reduce intra- and inter-individual differences. Wang et al. [40] analyzed electrode-frequency distribution maps (EFDMs) with CNNs, noting high-frequency bands' effectiveness in emotion recognition. Liu et al. [41] introduced CD-EmotionNet, a transfer learning-based model for enhancing emotion recognition with few-channel EEG data, marking a step towards cross-device adaptability.

III. METHODOLOGY

A. Overall Framework

This section introduces our Contrastive Learning based on the Diagonal Transformer Autoencoder (CLDTA). As illustrated in Fig. 1, the architecture encompasses both the pre-training procedure of contrastive learning and the calibration-prediction process in emotion recognition. The pre-training phase of CLDTA involves five key components: data preprocessing and feature extraction, augmentation, the DTA Encoder, the projector, and the contrastive loss function. Initially, samples are drawn from the EEG data bank and then processed and feature extracted followed by generating a broader sample range through the data augmentation module.

The DTA Encoder subsequently extracts emotion features based on brain networks from each EEG signal. Ultimately, the projector maps the properties into a high-dimensional feature space to compute the contrastive loss, optimizing the DTA Encoder and projector. During the calibration-prediction phase, the model, which integrates the pre-trained DTA Encoder and an initialized classifier, is fine-tuned using a small set of labeled samples from new subjects. This step enables accurate emotion detection in new subjects. Once calibrated, the model is then ready for emotion recognition tasks.

B. Data preprocess

The initial step in our process is to preprocess EEG signals to yield high-quality, artifact-free data. To obtain a more relevant and lower-dimensional representation for emotion recognition, we utilize the widely-used differential entropy (DE) feature, which is defined as follows:

$$DE(X) = - \int f(x) \log f(x) dx = \frac{1}{2} \log(2\pi e \sigma^2) \quad (1)$$

where σ^2 is the variance of the signal. Differential entropy features of each segment were extracted separately in the δ (0.1-4 Hz), θ (4-8 Hz), α (8-13 Hz), β (13-31 Hz), and γ (31-50 Hz) frequency bands. In one experiment of a subject, the DE features trained from continuous samples across time are concatenated and smoothed with a linear dynamic system (LDS) model [12].

C. Data Augmentation

Data augmentation enhances our model by diversifying data representation and acting as a regularizer to improve robustness and performance. We have adopted effective augmentation techniques for DE data, specifically MixUp method [42] and Masking technique [43], after thorough evaluation.

(1) MixUp

MixUp facilitates the model's ability to discern shared information among positive pairs of samples. The MixUp data augmentation process creates a new sample, by linearly combining a pair of randomly selected training samples, x_i and x_j as follows:

$$x' = \lambda' x_i + (1 - \lambda') x_j \quad (2)$$

where λ is a value sampled from a Beta distribution.

(2) Masking

The masking technique, otherwise referred to as channels dropout, has been demonstrated to yield superior results with sizable training sets [44]. This data augmentation method applies a mask that sets a random subset of channels to zero, introducing controlled noise and distortion. This procedure can be mathematically expressed as:

$$x'_i = x_i * \text{mask} \quad (3)$$

where 'mask' is a vector of zeroes and ones of length 62.

D. DTA Encoder

This section introduces the Diagonal Transformer Autoencoder (DTA), as depicted in Fig. 2. It draws on the fundamental principles of the Transformer encoder [32].

> TAFFC-2024-03-0186 <

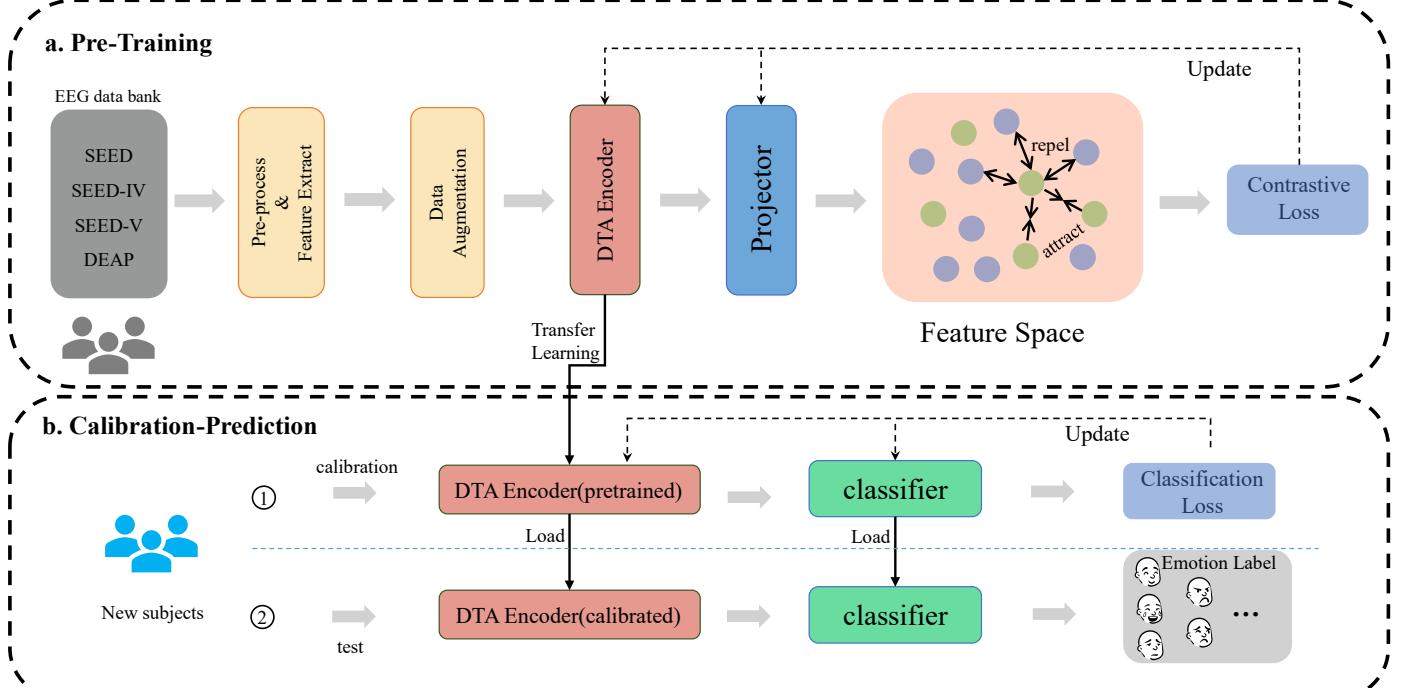


Fig. 1. Overview of the Transfer Learning Pipeline Using Contrastive Learning based on Diagonal Transformer Autoencoder (CLDTA). (a) Pre-Training: This phase involves preparing the EEG data from SEED, SEED-IV, SEED-V, and DEAP datasets, which undergoes preprocessing and data augmentation before being fed into the DTA Encoder. The encoder's output is then projected, and the model is updated based on contrastive loss, which aims to cluster similar emotion features closer in the feature space while pushing dissimilar ones apart, as indicated by the "attract" and "repel" arrows among subjects' representations. (b) Calibration-Prediction: This phase consists of two steps. First, a small subset of labeled samples from a new subject is collected to calibrate the pre-trained DTA Encoder and classifier. Next, the calibrated DTA Encoder and classifier are then utilized for subsequent emotion recognition in the same subject. Calibration adjusts the model to the new subject's EEG for better accuracy and the classifier links features to emotions for predicting the subject's emotional state.

Following the approach in [45], we incorporate a diagonal masking strategy (highlighted in blue in Fig. 2) to extract brain network knowledge, effectively bridging the gap between full-channel EEG data and fewer-channel EEG datasets. Moreover, we use an information separation mechanism (indicated by orange dashed lines in Fig. 2) to isolate the learned knowledge, thereby enhancing the model's interpretability.

(1) Diagonal Masking Strategy

The self-attention mechanism tends to assign excessively high attention weights to nodes themselves, as shown in Fig. 3(a). When processing EEG data with fewer number of channels, this sparsity of information can result in diminished accuracy, depicted in Fig. 3(b). To counteract this issue, Fig. 3(c) reveals that, in the pre-training stage, we capitalize on the Transformer's high parallel processing capability to focus on learning the brain network knowledge provided by full-channel EEG datasets through the Diagonal Masking strategy.

The attention mechanism of the Transformer consists of three vectors query, key and value (QKV). V is updated based on the matching degree of Q and K (i.e., attention matrix A). In this process, Diagonal Masking Operation is like $A_{ii} = 0$.

$$\text{Diagmask}(A) = (J - I) * A \quad (4)$$

where J represents a full matrix of ones, and I represents an identity matrix.

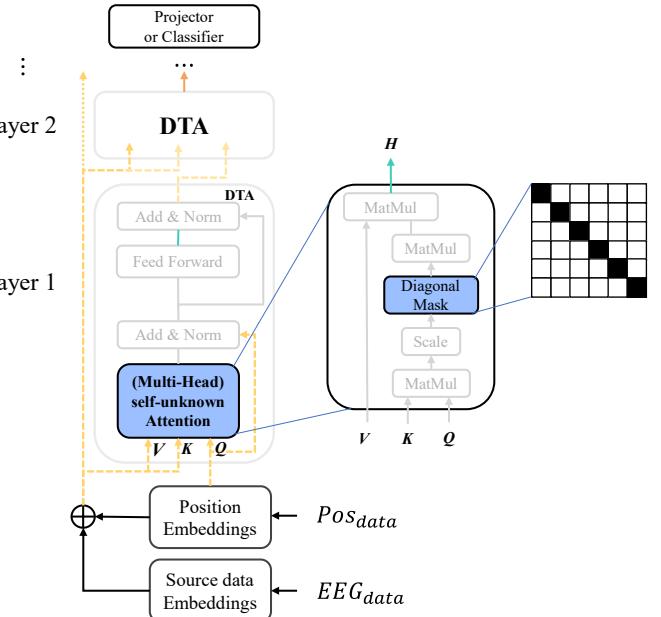


Fig. 2: Architecture of the DTA. The blue box signifies the diagonal masking strategy, and the orange dashed arrows represent the information separation mechanism, marking adaptations from the standard Transformer design.

> TAFFC-2024-03-0186 <

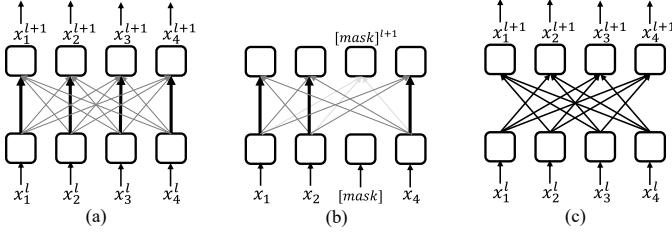


Fig. 3. Visualization of Self-Attention Weights and Diagonal Masking strategy in EEG Channel Analysis. Line thickness and shade indicate connection strength and influence. (a) Predominant self-attention in node weighting. (b) Self-Attention Weights after masking node 3 data. (c) Enhanced learning brain network knowledge through Diagonal Masking in pre-training.

(2) Information Separation Mechanism

The Transformer architecture uses residual connections, which impacts the Diagonal Masking Operation's self-unknown-attention abilities. Furthermore, the connections between EEG channels only can be understood by examining attention weights. An input separation method has been implemented to address this issue. Through this mechanism, the key value input (KV) for each encoding layer is isolated from the network flow and is fixed as a combination of the input encoding and position encoding. The query input (Q) is the only component that gets updated across layers. Fig. 2 illustrates this information isolation mechanism within the DTA, indicated by orange dashed arrows.

(3) Position Embedding and Source Data Embedding

The query (Q) input employs position embedding to transform the 3D coordinates of 62 nodes, derived from the 10-20 System, into the \$d_{model}\$ dimension using nonlinear mapping, thereby integrates prior spatial knowledge (\$P_{emb}\$) into the model. To augment the position encoding's expressiveness, we incorporate learnable position encoding (\$L_{emb}\$). Analogously, source data embedding encodes the differential entropy (DE) features to conform to the \$d_{model}\$ dimension size. The encoding formula is as follows:

$$P_{emb} = f_2(\text{acvtivate}(f_1(pos_{data}))) \quad (5)$$

$$S_{emb} = f_4(\text{acvtivate}(f_3(DE))) \quad (6)$$

$$Q^1 = P_{emb} + L_{emb} \quad (7)$$

$$K^1 = V^1 = Q^1 + S_{emb} \quad (8)$$

In the above formula, \$pos_{data}\$ represents the three-dimensional coordinates of the channel, \$P_{emb}\$ is the a priori position encoding, \$R_{emb}\$ is the learnable position embedding and \$S_{emb}\$ is the source data embedding. \$f_i(\cdot)\$ is the linear function, and \$\text{acvtivate}(\cdot)\$ is the activation function.

(4) Self-unknown Attention

As shown in Fig. 2, we have two encoding inputs, Source data Embedding and Position Embeddings, which are, respectively, \$X = [X_1, \dots, X_n]^T \in R^{n \times d_{model}}\$ and \$P = [P_1, \dots, P_n]^T \in R^{n \times d_{model}}\$, where \$d_{model}\$ is the feature dimension and \$n\$ is the number of channels. \$K\$ and \$V\$ are fixed in all layers, while the query \$Q^i\$ is updated with each layer.

Considering the input \$Q^i K^i V^i\$ of the i-th encoding layer and the output \$H^i\$, the formula is as follows:

$$H^i = \begin{cases} SUA(Q^i, K, V) = \text{Softmax}\left(\text{Diagmask}\left(\frac{Q^i K^T}{\sqrt{d}}\right)\right)V, & \text{if train} \\ SA(Q^i, K, V) = \text{Softmax}\left(\frac{Q^i K^T}{\sqrt{d}}\right)V & , \text{if test} \end{cases} \quad (9)$$

where \$SUA(\cdot)\$ represents the self-unknown attention layer, \$SA(\cdot)\$ represents the self attention layer and \$H = [H_1^i, \dots, H_m^i, \dots, H_n^i]\$.

In summary, while retaining the basic structure of the Transformer, the flow of information between the encoding layers in the CLDTA encoding module is as follows:

$$g(x) = \text{Norm}(\text{Add}(x, \text{FeedForward}(x))) \quad (10)$$

$$Q^{i+1} = DTA(Q^i, K, V) = g(\text{norm}(\text{Add}(Q^i, H^i))) \quad (11)$$

Here, we note that in the inference training process of CLDTA, the i-th element represented by \$Q_i\$ will not directly see the corresponding encoding representation from \$Q_i^1 = P^1\$ in any layer. However, during the testing phase, the diagonal masking mechanism is shut down, restoring it to self-attention.

E. The Projector

The nonlinear projector can help the basic encoder better learn the representation of downstream prediction tasks[46]. Here, we only use the Multilayer Perceptron (MLP), the formula is as follows:

$$Z = \text{Projector}(Q) \quad (12)$$

As shown in Fig. 4, the Projector mainly includes three fully connected layers with the number of hidden units decreasing sequentially from 128, 256, to 128. The corresponding positions in the figure show Batch Normalization, ELU and Dropout.

F. The Contrastive Loss

To measure the similarity of emotion-related features between two sets of samples, we can calculate the cosine similarity of the encoded representation vectors. The input batch samples \$G^A = [G_1^A, \dots, G_n^A]\$ and \$G^B = [G_1^B, \dots, G_n^B]\$ are transformed into \$Z^A\$ and \$Z^B\$ through the DTA encoder and the projector, respectively. Then, we can compute the cosine similarity of the feature sets between \$Z^A\$ and \$Z^B\$:

$$s(z_i^A, z_i^B) = \frac{z_i^A \cdot z_i^B}{\|z_i^A\| \|z_i^B\|}, s(z_i^A, z_i^B) \in [0, 1] \quad (13)$$

The purpose of contrastive loss is to maximize the similarity of the EEG signals within the positive pair as fully as possible. We adopt the normalized temperature-scaled binary cross-entropy with logits loss computed by

$$x_i = \frac{s(z_i^A, z_i^B)}{\tau} \quad (14)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (15)$$

$$l_n = -[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log (1 - \sigma(x_n))] \quad (16)$$

$$\text{loss} = \frac{1}{N} (l_1, \dots, l_N) \quad (17)$$

where \$\tau\$ is the temperature parameter for softmax. The variable \$y\$ can take the values of 0 or 1 and \$\sigma(x)\$ denotes the sigmoid function.

> TAFFC-2024-03-0186 <

The smaller the loss function, the more similar the samples in the same category and the more dissimilar the samples in different categories. Adopting this loss function allows a sample to be similar to multiple samples at the same time, thereby accelerating the training.

G. Calibration-prediction Process

In the calibration-prediction process, we use the pre-trained DTA encoder to extract emotional features and predict emotional labels from the representations. We optimize the parameters of the pre-trained model and classifier using the cross-entropy loss function.

The classifier is utilized to predict emotional labels from the representations extracted from the DTA encoder. As depicted in Fig. 4, the classifier primarily comprises two fully connected layers.

$$\text{Label} = \text{Classifier}(\text{DTA}_{\text{pretrained}}(\text{DE}, \text{pos}_{\text{data}})) \quad (18)$$

Finally, when the loss function converges, it can be used for subsequent emotional recognition of the subjects.

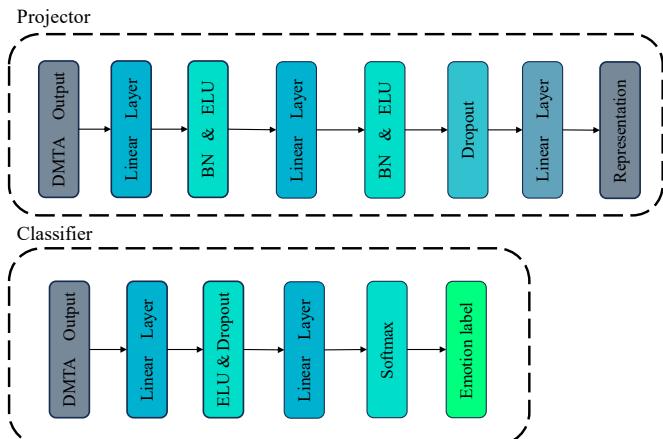


Fig. 4. The architecture of the Projector and Classifier. BN represent Batch Normalization. Linear layer represents fully-connected layer. ELU represents Exponential Linear Unit.

IV. EXPERIMENTS

In this section, we outline the datasets employed, elucidate the data preprocessing procedures, and expound upon the implementation details of the model. Subsequently, We define the evaluation procedures and introduce the advanced deep learning benchmarks used for comparison. Lastly, we discuss the methodologies deployed for analyzing the performance of our model.

A. Dataset

We first outline the datasets selected for this study and the rationale behind their selection:

- (1) SEED Dataset: Developed by Zheng and Lu [12], this dataset includes EEG data from 15 subjects who watched 15 Chinese film clips, eliciting three emotions: positive, negative, and neutral. Each subject participated in three sessions, watching one clip per session for a total of 15 trials.
- (2) SEED-IV Dataset: Introduced in [27], this dataset features EEG and eye movement data from 15 subjects (7 men and 8

women) responding to 72 film scenes depicting four emotions: joy, sorrow, neutrality, and anxiety. Subjects participated in three sessions with 24 trials each at different times.

(3) SEED-V Dataset: First utilized in [28], it comprises EEG and eye movement signals related to five emotions (happiness, sadness, neutral, fear, and disgust) from 15 film clips, with 16 subjects (6 males, 10 females) participating in three sessions.

(4) DEAP Dataset: Established by Koelstra et al. [13], this dataset consists of EEG signals from 32 channels and peripheral physiological signals from 8 channels, collected from 32 participants watching 40 one-minute music videos. Participants rated the videos on arousal, appeal, likes/dislikes, dominance, and familiarity.

The SEED series dataset is expected to be an excellent benchmark for pre-training the CLDTA model, as it features a significantly larger number of subjects compared to most publicly available datasets. The datasets were collected in controlled environments to induce specific emotions using video clips, with data captured via a 62-channel ESI NeuroScan system aligned with the International 10-20 system. They offer a broad range of emotional labels for a discrete emotional modeling approach, as opposed to a valence-arousal spectrum. Contrastingly, the DEAP dataset, with its different EEG equipment, data specifications, emotional stimuli, and labeling approach, presents unique challenges for cross-dataset classification tasks. This makes it an ideal candidate for assessing the model's performance across diverse datasets.

B. Data Preprocessing

To ensure data consistency, we re-processed the original EEG data from the datasets. This study primarily utilized the EEGLAB toolbox [47] in MATLAB for pre-processing, which includes data input, electrode positioning, filtering, baseline correction, manual identification and removal of bad segments and channels, independent component analysis(ICA), manual exclusion of irrelevant components, and re-referencing. For the SEED, SEED-IV, and SEED-V datasets, we initially applied a band-pass filter from 0.01 to 48 Hz and a 50 Hz notch filter to eliminate noise. The criteria for rejecting bad channels are as follows: channels with a flatline duration exceeding 5 seconds; channels whose variance exceeds 4 times the standard deviation of the total channel signal; and spatially adjacent channels with a correlation less than 0.6. The criteria for rejecting time segments are: if the variance in each time window exceeds 7 times the variance of the current channel, the window is discarded. EEGLAB's 'spherical' interpolation algorithm is employed to interpolate channels discarded due to volume conduction effects, assigning different interpolation weights based on the proximity of surrounding nodes. ICA is subsequently applied to remove artifacts likely caused by eye movements, muscle movements, or other environmental noise, with up to 5 ICA components being removed. The data is re-referenced using a sample mean reference. We utilize the last 30 seconds of each trial to ensure the stimulated emotions are sufficiently coherent and intense.

> TAFFC-2024-03-0186 <

For the DEAP dataset, we employed the same data preprocessing method. The data was first adjusted to match the 62-channel format of the SEED-series datasets, and missing channel data was filled with zeros. We adhered to the partitioning strategy outlined in [48] and [41], which converts the dataset into binary emotion recognition tasks by segmenting the valence dimension into positive/negative and the arousal dimension into high/low arousal, with the threshold for both dimensions set at 5. Thus, the processed data can be summarized as shown in TABLE I.

TABLE I
Summary of Preprocessed Dataset Details

Dataset	Subject	Session	Trial	Sample	Total
SEED	15	3	15	30	20250
SEED-IV	15	3	24	30	32400
SEED-V	15	3	15	30	21150
DEAP	32	1	40	30	38400

Note: Trial: refers to the number of trials selected. Sample: represents the number of samples per each trial. Total: signifies the total sample count for each dataset. For the SEED-V dataset, the first trial of data from subject 5 is missing.

C. Training Details

We trained our CLDTA on NVIDIA RTX 3080ti GPU, pretraining the model on SEED, IV, and V datasets. The CLDTA was configured to 4 layers, model dimension (d_{model}) to 32, hidden layer dimension to 64, and multi-head attention count to 4. The Projector flattens the data and maps it to 128 dimensions. The temperature hyperparameter τ for contrastive learning was set to 0.5.

For optimizing the contrastive learning model, we used the Adam optimizer [49], with the initial learning rate set to 1e-4, and weight decay set to 0.005 based on empirical standards. A random seed of 42 was set, batch size was configured to 256, epoch was set to 30, dropout was set to 0.1, and activation function was set as Exponential Linear Units (ELU) [50].

For the calibrating and transfer process of emotion recognition in MLP classifier, we used two hidden layers, each with 32 units. ELU were used between every two layers. We used cross-entropy loss and Adam optimizer for parameter optimization. The learning rate was empirically set to 1e-5. Batch size was empirically set to 128. We trained for 100 epochs with early stopping (maximal tolerance of 20 epochs without validation accuracy increase).

D. Test and Validation

We applied the leave-one-subject-out cross-validation (LOSOCV) method to assess our approach. In LOSOCV, each subject's data is alternately used for transfer learning, with the rest for training. For each test, an equal number of labeled samples per category is selected from the target subject's test set, excluding all other unlabeled samples from training. This process repeats for all subjects' data.

Subject-dependent experiments use a small set of labeled samples from target subjects for transfer learning, with the

remaining data for accuracy testing. The training and testing set division follows protocols from [41] and [12]. For SEED, training involves the first 9 trials per session, with the next 6 trials for testing. SEED-IV uses the first 16 trials for training and the last 8 for testing. SEED-V employs a triple cross-validation (10 for training, 5 for testing) for five emotion tasks. DEAP uses an 80% training and 20% testing split per subject.

In strictly subject-independent experiments, when no target subject calibration samples are available, calibration uses source subjects' data, followed by testing on the target subjects.

E. Performance Comparison

To investigate the effectiveness of our contrast learning method, we compared it with several notable emotion recognition methods, including A-LSTM [51], DGCNN [17], BiDANN [23], SSL-EEG [52], RGNN [24], GMSS [15], and PR-PL [25]. These methods are emblematic of prior research in emotion recognition. Their results were either directly quoted or replicated from the literature to ensure a reliable comparison with our proposed method. It's important to note that our results are compared only with advanced models under the same standard experimental settings. In our performance comparison protocols, results reproduced by our team are marked with an asterisk (*).

F. Methods for Analyzing Model Performance

(1) Model Stability and Channel Reduction

In the SEED series dataset, with its 62 channels from various brain locations, the excessive number of channels not only raises computational demands but also hampers the practicality of aBCI systems. Hence, it's essential to minimize channel use while analyzing EEG data. Our model calibration tests involved randomly masking EEG channels to assess the impact of channel quantity on recognition accuracy.

(2) Identifying Brain Regions for Emotion Recognition

EEG channels correspond to different brain cortex areas, each associated with specific physiological functions. To pinpoint crucial regions or channels for emotion recognition, we analyzed location encoding data. Calculating the cosine similarity between channels helped us identify the importance of nodes and their community groupings.

(3) Contrastive Learning Evaluation

We evaluated the impact of contrastive learning by visualizing features before and after encoding and by measuring inter-class divergence (ICD) and intra-class similarity (ICS). ICD evaluates the similarity level among samples of the same class in the embedding space, while ICS assesses the separation degree between different class samples. A smaller intra-class distance implies higher intra-class similarity, and a larger inter-class distance indicates greater separation.

$$CD = E[\|f(x) - f(y)\|_2^\alpha], \alpha > 0 \text{ and } (x, y) \sim p_{pos} \quad (19)$$

$$ICS = E[\|f(x) - f(y)\|_2^\alpha], \alpha > 0 \text{ and } (x, y) \sim p_{neg} \quad (20)$$

In these calculations, p_{pos} denotes scenarios where the labels of the pair are matching, while p_{neg} refers to scenarios where the labels do not match.

V. RESULT AND DISCUSSION

Drawing from the analysis presented in Section IV, Part A, we first evaluate the CLDTA model's performance in subject-dependent and subject-independent setups on the SEED series datasets. Then, we evaluate the cross-device and cross-dataset classification tasks on the DEAP dataset.

A. Emotion Recognition Performance on the SEED series Dataset

(1) Subject-dependent Evaluation

Three configurations of DTA were tested: DTA without transfer learning (DTA w/o TF), DTA with transfer learning within the same dataset (DTA-Single-Dataset), and DTA with transfer learning across multiple datasets (DTA-Multi-Dataset). The experimental results are shown in Table II. The results underscore the benefits of transfer learning, especially when applied across datasets, in improving the model's effectiveness. In subject-dependent evaluations, DTA shows competitive or superior performance compared to advanced models like GMSS, achieving an accuracy of 95.09% on the SEED dataset and the highest accuracy on SEED-IV and SEED-V, indicating its capability to learn stable subject features.

(2) Subject-independent Evaluation

In the subject-independent experiments detailed in Table III, it is evident that the CLDTA model outperforms the SVM baseline by achieving respective performance enhancements of 32.4%, 23.8%, and 29.9% on the SEED, SEED-IV, and SEED-V datasets. Furthermore, the CLDTA model attains state-of-the-art results on SEED-IV and SEED-V, with accuracies of 64.11% and 61.45%, respectively.

Furthermore, CLDTA consistently presents a notably low accuracy standard deviation in both testing scenarios, demonstrating its strong discrimination and generalization abilities. This comprehensive performance across different testing conditions confirms the effectiveness of the proposed transfer learning strategy in optimizing network performance, highlighting CLDTA as a viable approach for practical emotion recognition applications. However, its performance on the SEED dataset did not reach the most advanced level, which may be attributed to the dataset's broad emotional categories (positive, negative, neutral) as opposed to the more granular labels found in SEED-IV and SEED-V. These findings suggest that the efficacy of the model's learning is influenced by the granularity of emotion labeling.

TABLE II

EEG Emotion Recognition: Comparison with State-of-the-Art Methods on SEED, SEED-IV, SEED-V (Mean and SD%)

MODEL	Dataset					
	SEED		SEED-IV		SEED-V	
	Acc.	Std.	Acc.	Std.	Acc.	Std.
SVM[53]	83.99	9.27	56.61	20.05	69.5	10.28
A-LSTM[51]	88.61	10.16	69.50	15.65	-	-
DGCNN[17]	90.4	8.49	65.97	15.03	-	-
BiDANN[23]	92.38	7.04	70.29	12.63	-	-
SSL-EEG[52]	83.32	9.20	63.59	19.82	-	-
RGNN[24]	94.24	5.95	79.37	10.54	-	-
GMSS[15]	96.48	4.63	86.37	11.45	-	-
PR-PL[25]	94.84	9.16	83.33	10.61	-	-
DTA w/o TF	90.44	8.49	81.88	13.29	77.92	11.17
DTA-Single-Dataset	93.12	5.02	82.12	6.52	78.33	9.61
DTA-Multi-Dataset	95.09	4.48	88.3	4.62	80.15	8.33

— indicates the experiment results are not reported on that dataset.

Table III
Subject-Independent Accuracies on SEED, SEED-IV, SEED-V (Mean and SD%)

MODEL	Dataset					
	SEED		SEED-IV		SEED-V	
	Acc.	Std.	Acc.	Std.	Acc.	Std.
SVM[53]	56.73	16.29	51.78	12.85	47.3	16.53
A-LSTM[51]	72.18	10.85	55.03	09.28	-	-
DGCNN[17]	79.95	09.02	52.82	09.23	-	-
SSL-EEG[52]	67.52	12.73	53.62	08.47	-	-
GMSS[15]	76.04	11.91	62.13	08.33	-	-
CLDTA	75.09	05.88	64.11	04.62	61.45	10.82

— indicates the experiment results are not reported on that dataset.

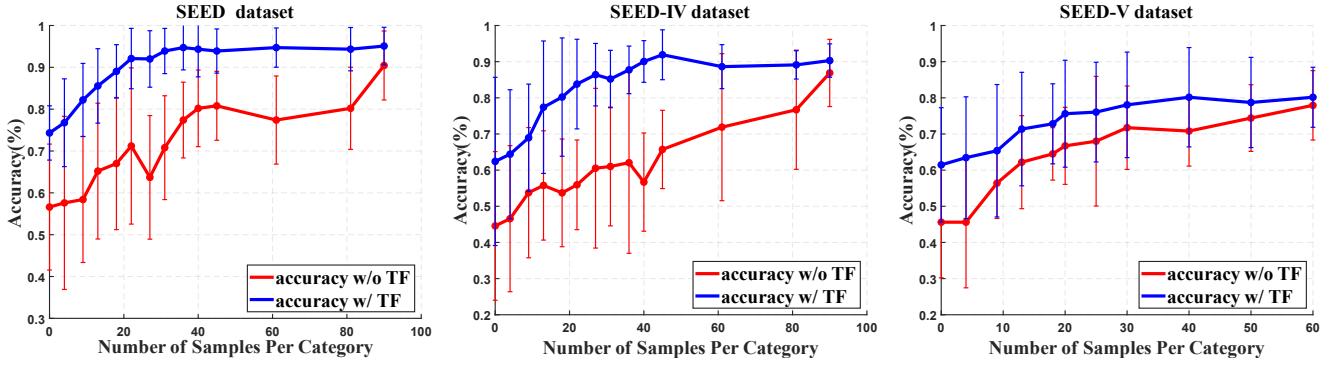


Fig. 5. Performance of Fully-Supervised vs. Transfer-Learning-Based CLDTA Models in Calibration Tests Across SEED Datasets: (a) SEED, (b) SEED-IV, (c) SEED-V

B. Cross-device Cross-electrode Evaluation on the DEAP Dataset

We compared two transfer learning strategies: one employing a model pre-trained on the SEED series datasets (SEED2DEAP) and another pre-trained on DEAP itself (DEAP2DEAP), against a baseline model with no transfer learning (Rand2DEAP). This comparison elucidates the impact of transfer learning on model efficacy in varying experimental setups. The results, as shown in Table IV, indicate that our CLDTA model, leveraging transfer learning, achieved superior accuracy. Specifically, SEED2DEAP excelled in arousal classification with a 94.11% accuracy rate and a 2.1% standard deviation, while DEAP2DEAP showed higher accuracy in Valence classification at 94.58% with a 1.4% standard deviation. These results affirm the effectiveness of our proposed transfer learning strategy in capturing cross-device and cross-electrode EEG emotion features.

TABLE IV
CLDTA vs. State-of-the-Art Methods on DEAP: Valence and Arousal Classification Accuracies

Methods	Accuracy (Mean / SD) (%)	
	Valence	Arousal
SVM[53]	72.59 / 9.73	74.44 / 9.84
CD-EmotionNet [41]	86.29 / 9.71	84.16 / 10.86
DGCNN[17]	86.32 / 6.04	83.68 / 5.68
Rand2DEAP	81.92 / 3.53	83.37 / 4.33
SEED2DEAP	93.31 / 1.80	94.11 / 2.10
DEAP2DEAP	94.58 / 1.40	92.58 / 1.80

C. Calibration Test

To assess performance with limited labeled samples, we explored how different quantities of labeled samples affect model calibration, comparing models with and without transfer learning. Figure 5 illustrates that when employing different quantities of calibration samples for fine-tuning, the accuracy of the CLDTA model markedly surpasses that of the fully-supervised baseline across the entire range of sample

sizes, with the most pronounced advantage observed in scenarios with limited labeled data. Specifically, for the SEED dataset, CLDTA's performance nearly matches full-supervised training (90.44%) with over 20 labeled samples per category. For SEED-IV, CLDTA reaches 87.88% of full-supervised training with more than 32 labeled samples per category. For SEED-V, CLDTA's performance is close to full-supervised training (77.92%) with over 13 labeled samples per category. Beyond 40 calibration labels per category, the performance of all pre-training models converges.

In addition, we also recorded the time consumed by the model during the calibration prediction phase, as shown in Table V. The number of training iterations required for calibrating the pre-trained model is 17% of that required by the randomly initialized model. In terms of training time, this represents a time saving of 91.48%. This demonstrates that pre-trained models are both faster and more stable in calibration compared to fully-supervised models.

For the DEAP dataset, calibration tests were conducted on three models: SEED2DEAP, DEAP2DEAP, and Rand2DEAP. Results shown in Fig. 6 indicate that pre-trained models on SEED and DEAP achieve nearly similar performances with limited samples, with a mean accuracy difference of 3.7%. This indicates that CLDTA effectively captures subject-invariant emotional traits, showing resilience to differences in channel numbers and device types. The baseline model showed a higher tendency for overfitting compared to the pre-trained models, which adapted better.

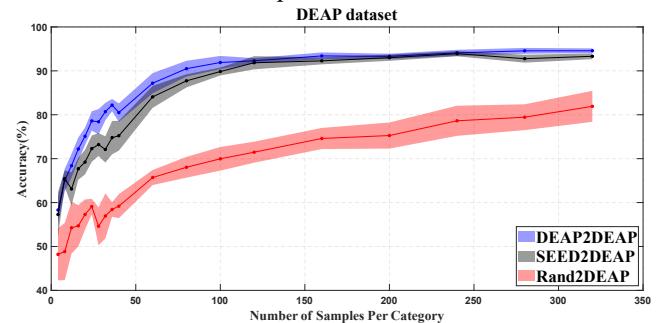


Fig. 6. Calibration Test Results on DEAP Dataset: Comparing SEED2DEAP, DEAP2DEAP, and Rand2DEAP Model Performances.

TABLE V
Comparison of Calibration Training Convergence between
Pre-trained and Randomly Initialized Models

Model	Epochs	Times(seconds)
Randomly Initialized DTA	1084	636.28
Pre-trained DTA	187	54.23

D. Stability Analysis

For practical applications, minimizing the number of electrodes is advantageous for both feasibility and user comfort. Challenges including disconnections due to head movements, short circuits from excessive conductive paste, and potential electrode malfunctions can impair model performance. To evaluate our model's resilience in the face of such issues, we simulated real-world conditions such as electrode failure and noise interference.

We used the model pre-trained on the SEED series dataset for our experiments. In the electrode failure test, we simulated failures by setting channel data to zero or replacing it with data from nearby channels, with the number of failed channels ranging from 1 to 40. In the noise interference test, we added Gaussian noise with intensity varying from 0.1 to 3 times the sample variance. The results, shown in Fig.7, indicate the pre-trained model's superior anti-interference capability compared to a fully supervised model. A small number of electrode failures slightly improved performance by 1.21%, suggesting that redundant channels may introduce noise in emotion recognition tasks. Performance declines in the pre-trained model when failures exceed 26 electrodes, while the fully supervised model's performance gradually decreases with more failures.

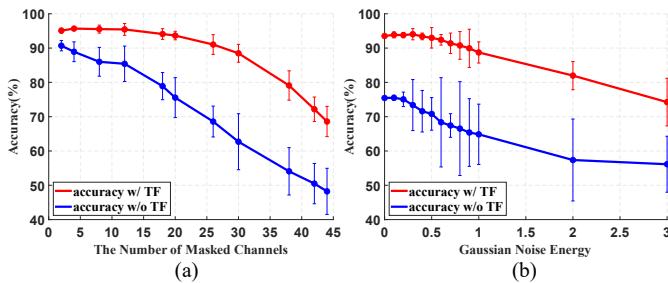


Fig. 7. Evaluating Pre-trained and Fully-Supervised Models' Resilience to Electrode Failure and Noise Interference. (a) Electrode failure experiment, (b) Noise interference experiments.

During the noise interference experiment, the pre-trained model consistently outperformed the fully supervised model at low noise levels. When the noise intensity was under 1, the transfer model's performance decreased by only 4%, a minor reduction compared to the 10% drop in the fully supervised model. However, as noise intensity increased from 1 to 3 times the variance, the accuracy of both models dropped—the pre-trained model by 14.5% and the fully-supervised model by

8.7%. The pre-trained model's initial stability may be due to its reliance on sophisticated features learned during pre-training, making it more resistant to minor disturbances. Yet, high noise levels impact the pre-trained model more as it may inaccurately associate enhanced noise with previously learned features, leading to performance drops. Conversely, the fully-supervised model adapts better to high noise levels, possibly because it continuously fine-tunes parameters to accommodate all variations, including noise.

E. Explainability and Connectivity analysis

To investigate the role of different brain regions in emotion recognition, we conducted a connection analysis after the model stabilized, focusing on the 10-20 system. By computing cosine similarity between node positions to form an adjacency matrix and retaining only connections exceeding the mean plus 1.8 standard deviations, our analysis (Fig. 8) highlights significant involvement of the frontal and temporal lobes in emotion processing, along with observed asymmetry in brain hemisphere activities. These findings align with previous studies [52],[54],[55], suggesting a correlation with the spatial distribution of emotions and activation of frontal-parietal networks in response to emotional stimuli. This underscores the distinct EEG signal characteristics during emotion recognition and suggests potential for future research using advanced graph theory to further elucidate the complex interactions between brain regions and emotions.

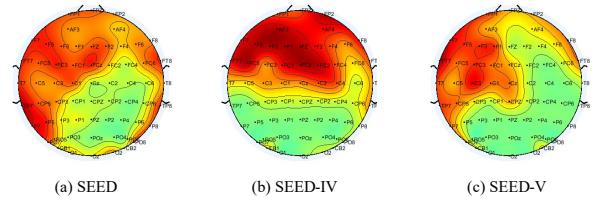


Fig. 8. Degree centrality of brain connectivity learned by the CLDTA, with darker colors indicating stronger connections to other regions.

F. Feature Space Visualization

This section describes the extraction of features by the CLDTA model that align between subjects, particularly when encountering new participants. We utilize the t-SNE algorithm to evaluate the model's performance on previously unseen subjects. Due to space limitations and for clarity, we randomly chose three participants and visualized their spatial characteristics for both positive and negative emotions. Figure 9 illustrates the feature distributions of these three subjects from the SEED dataset in a two-dimensional space using t-SNE.

We analyzed the ICD and ICS metrics, as detailed in equations (19) and (20), with results displayed in Table VI. Initial observations from Fig. 9(a) and (b) highlight subject variability; the same emotions across different subjects are widely spaced, often overlapping with different emotions from other subjects, underscoring the challenge of cross-subject

> TAFFC-2024-03-0186 <

recognition. Conversely, Fig. 9(c) shows that the CLDTA model effectively blends data across subjects while maintaining distinct emotion categories. This result indicates a substantial reduction in subject-specific features, as the model projects subject features into an emotion-centric space that is independent of the individual without needing subject-specific calibration. It suggests that pre-training via contrastive learning effectively reduces subject variability while maintaining the distinction of emotions, thereby enabling cross-subject emotion recognition. Post-calibration, as depicted in Fig. 9(d), both metrics exhibit further enhancement, demonstrating the model's capability to quickly adjust to new subjects with minimal labels, thereby markedly improving both performance and user experience.

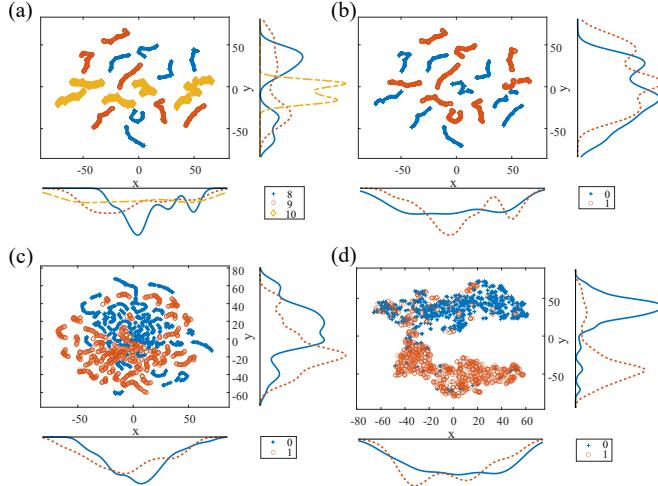


Fig. 9. Visualization of latent features using t-SNE on the SEED dataset. (a) t-SNE results of original differential entropy (DE) features for subjects 8, 9, 10, colored according to different subjects. (b) Original features represented through color-coding for positive and negative emotions. (c) Features extracted by the pre-trained CLDTA model without calibration for new subjects. (d) Enhanced performance of the model following subject calibration using 20 labeled samples.

TABLE VI

Computed inter-class divergence (ICD) and intra-class similarity (ICS) Metrics

	ICD	ICS
before encode: (a), (b)	9.876	0.1127
after encode: (c)	37.49	0.0569
after calibration: (d)	73.75	0.0323

G. Ablation study

To examine the role of data augmentation in enhancing our model's EEG emotion recognition capability, we conducted an ablation study with the CLDTA model. This study assessed the effects of data augmentation techniques on model performance, as detailed in Table VII. Utilizing a singular augmentation approach, CLDTA with Masking (CLDTA w/

Mask) demonstrated superior results, underscoring the mask's significant contribution to improving EEG emotion signal discernibility. While CLDTA with MixUp (CLDTA w/ MixUp) didn't achieve as high accuracy as masking, it exhibited lower variance, suggesting MixUp's effectiveness in enhancing sample continuity and adaptability to new subjects. Notably, combining both augmentation methods further increased accuracy and reduced variance, indicating their complementary benefits in aiding the model's learning of distinct emotional representations.

TABLE VII
Ablation study: subject-dependent classification accuracy (mean/std) on SEED, SEED-IV, and SEED-V

Method	Accuracy (Mean / SD) (%)		
	SEED	SEED-IV	SEED-V
CLDTA w/ MixUp	88.21/5.27	83.65/8.05	76.50/8.28
CLDTA w/ Mask	92.50/7.63	84.43/9.57	78.80/9.65
CLDTA w/ both	95.09/4.48	88.30/4.62	80.15/8.3

VI. CONCLUSIONS

This paper introduces a Transfer Learning framework utilizing contrastively pre-trained CLDTA, which encode EEG signals into subject-independent emotional representations, regardless of channel count. We tested our model against four prominent emotional databases, SEED, SEED-IV, SEED-V, and DEAP, comparing it with current benchmarks. Our CLDTA model presents several advantages over existing emotion recognition methods. It dynamically leverages spatial characteristics of EEG channels based on the 10-20 system, enabling it to accommodate diverse emotion datasets with varying channel counts. Through contrastive learning, the model potentially uncovers shared temporal-spatial patterns among different emotion categories, offering insights with neurophysiological significance. Moreover, CLDTA's ability to model new subjects with fewer calibration data and its enhanced anti-interference capabilities reduce the reliance on costly label collection and manual feature extraction. This facilitates broader and quicker deployment of emotion recognition systems, improving their practical applicability.

However, the primary training data source is the SEED series, and despite employing multiple data augmentation techniques, the limited diversity could impact the robustness and generalizability of aBCI models in real-world applications. Additionally, further discussion and research on the topic of negative transfer remain necessary. Lastly, the practical deployment of the model, particularly in environments with a lower signal-to-noise ratio than laboratory conditions, has yet to be tested.

To enhance applicability in real-world scenarios, future work will aim to collect a more diverse dataset covering a wider range of ages and scenarios, explore the feasibility of large-scale emotional BCI models, and plan to achieve high

> TAFFC-2024-03-0186 <

performance with fewer EEG channels. It is only when aBCI can provide stable and effective performance across sessions, subjects, and dataset tasks that it can be expected to manage the complex and varied emotional recognition scenarios in real-life.

ACKNOWLEDGMENT

The authors extend their appreciation to Xiangkai Qiu and Shenglin Wang for their invaluable input on manuscript composition and figure design. Gratitude is also due to Professor Bao-Liang Lu's team and the BCMI Laboratory for their swift provision of the SEED-related dataset and code.

REFERENCES

- [1] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, Jul. 2020, doi: 10.1016/j.inffus.2020.01.011.
- [2] Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing Deep Learning Towards Multi-Modal Bio-Sensing and Vision-Based Affective Computing," *IEEE Trans. Affective Comput.*, vol. 13, no. 1, pp. 96–107, Jan. 2022, doi: 10.1109/TAFFC.2019.2916015.
- [3] X. Hu, J. Chen, F. Wang, and D. Zhang, "Ten challenges for EEG-based affective computing," *Brain Science Advances*, vol. 5, no. 1, pp. 1–20, Mar. 2019, doi: 10.1177/2096595819896200.
- [4] D. Wu, B.-L. Lu, B. Hu, and Z. Zeng, "Affective Brain–Computer Interfaces (aBCIs): A Tutorial," *Proc. IEEE*, vol. 111, no. 10, pp. 1314–1332, Oct. 2023, doi: 10.1109/JPROC.2023.3277471.
- [5] K. Kamble and J. Sengupta, "A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals," *Multimed Tools Appl.*, vol. 82, no. 18, pp. 27269–27304, Jul. 2023, doi: 10.1007/s11042-023-14489-9.
- [6] J. Cai, R. Xiao, W. Cui, S. Zhang, and G. Liu, "Application of Electroencephalography-Based Machine Learning in Emotion Recognition: A Review," *Front. Syst. Neurosci.*, vol. 15, p. 729707, Nov. 2021, doi: 10.3389/fnsys.2021.729707.
- [7] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Information Fusion*, vol. 102, p. 102019, Feb. 2024, doi: 10.1016/j.inffus.2023.102019.
- [8] Y. Zhang, S. Yang, G. Cauwenberghs, and T.-P. Jung, "From Word Embedding to Reading Embedding Using Large Language Model, EEG and Eye-tracking," arXiv, Jan. 28, 2024. Accessed: Mar. 18, 2024. [Online]. Available: <http://arxiv.org/abs/2401.15681>
- [9] S. Katsigianis 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 Inform.*, vol. 22, no. 1, pp. 98–107, Jan. 2018, doi: 10.1109/JBHI.2017.2688239.
- [10] P. Lakhan *et al.*, "Consumer Grade Brain Sensing for Emotion Recognition," *IEEE Sensors J.*, vol. 19, no. 21, pp. 9896–9907, Nov. 2019, doi: 10.1109/JSEN.2019.2928781.
- [11] Z. Wan, J. Huang, H. Zhang, H. Zhou, J. Yang, and N. Zhong, "HybridEEGNet: A Convolutional Neural Network for EEG Feature Learning and Depression Discrimination," *IEEE Access*, vol. 8, pp. 30332–30342, 2020, doi: 10.1109/ACCESS.2020.2971656.
- [12] Wei-Long Zheng and Bao-Liang Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," *IEEE Trans. Auton. Mental Dev.*, vol. 7, no. 3, pp. 162–175, Sep. 2015, doi: 10.1109/TAMD.2015.2431497.
- [13] S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.
- [14] C. Li, Y. Hou, R. Song, J. Cheng, Y. Liu, and X. Chen, "Multi-channel EEG-based emotion recognition in the presence of noisy labels," *Sci. China Inf. Sci.*, vol. 65, no. 4, p. 140405, Apr. 2022, doi: 10.1007/s11432-021-3439-2.
- [15] Y. Li *et al.*, "GMSS: Graph-Based Multi-Task Self-Supervised Learning for EEG Emotion Recognition," *IEEE Trans. Affective Comput.*, vol. 14, no. 3, pp. 2512–2525, Jul. 2023, doi: 10.1109/TAFFC.2022.3170428.
- [16] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, San Diego, CA, USA: IEEE, Nov. 2013, pp. 81–84, doi: 10.1109/NER.2013.6695876.
- [17] 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, Jul. 2020, doi: 10.1109/TAFFC.2018.2817622.
- [18] A. Iyer, S. S. Das, R. Teotia, S. Maheshwari, and R. R. Sharma, "CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings," *Multimed Tools Appl.*, vol. 82, no. 4, pp. 4883–4896, Feb. 2023, doi: 10.1007/s11042-022-12310-7.
- [19] A. Sakalle, P. Tomar, H. Bhardwaj, D. Acharya, and A. Bhardwaj, "A LSTM based deep learning network for recognizing emotions using wireless brainwave driven system," *Expert Systems with Applications*, vol. 173, p. 114516, Jul. 2021, doi: 10.1016/j.eswa.2020.114516.
- [20] Z. Li, G. Zhang, L. Wang, J. Wei, and J. Dang, "Emotion recognition using spatial-temporal EEG features through convolutional graph attention network," *J. Neural Eng.*, vol. 20, no. 1, p. 016046, Feb. 2023, doi: 10.1088/1741-2552/acb79e.
- [21] Y. M. Chi, T.-P. Jung, and G. Cauwenberghs, "Dry-Contact and Noncontact Biopotential Electrodes: Methodological Review," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 106–119, 2010, doi: 10.1109/RBME.2010.2084078.
- [22] 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, Jul. 2023, doi: 10.1109/TAFFC.2022.3164516.
- [23] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong, "A Novel Neural Network Model based on Cerebral Hemispheric Asymmetry for EEG Emotion Recognition," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, pp. 1561–1567, doi: 10.24963/ijcai.2018/216.
- [24] P. Zhong, D. Wang, and C. Miao, "EEG-Based Emotion Recognition Using Regularized Graph Neural Networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2022, doi: 10.1109/TAFFC.2020.2994159.
- [25] R. Zhou *et al.*, "PR-PL: A Novel Prototypical Representation Based Pairwise Learning Framework for Emotion Recognition Using EEG Signals," *IEEE Trans. Affective Comput.*, pp. 1–14, 2023, doi: 10.1109/TAFFC.2023.3288118.
- [26] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, in IJCAI'16. New York, New York, USA: AAAI Press, 2016, pp. 2732–2738.
- [27] 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, doi: 10.1109/TCYB.2018.2797176.
- [28] 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. Dev. Syst.*, vol. 14, no. 2, pp. 715–729, Jun. 2022, doi: 10.1109/TCDS.2021.3071170.
- [29] Wei-Long Zheng, Bo-Nan Dong, and Bao-Liang Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL: IEEE, Aug. 2014, pp. 5040–5043, doi: 10.1109/EMBC.2014.6944757.
- [30] X. Wang, Y. Ma, J. Cammon, F. Fang, Y. Gao, and Y. Zhang, "Self-Supervised EEG Emotion Recognition Models Based on CNN," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1952–1962, 2023, doi: 10.1109/TNSRE.2023.3263570.
- [31] 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*, Nice France: ACM, Oct. 2019, pp. 176–183, doi: 10.1145/3343031.3350871.
- [32] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Aug. 01, 2023. Accessed: Dec. 25, 2023. [Online]. Available: <http://arxiv.org/abs/1706.03762>

> TAFFC-2024-03-0186 <

- [33] 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, doi: 10.1109/JSEN.2022.3144317.
- [34] A. Kondacs and M. Szabó, "Long-term intra-individual variability of the background EEG in normals," *Clinical Neurophysiology*, vol. 110, no. 10, pp. 1708–1716, Oct. 1999, doi: 10.1016/S1388-2457(99)00122-4.
- [35] C.-D. Wang *et al.*, "Cross-Subject Tinnitus Diagnosis Based on Multi-Band EEG Contrastive Representation Learning," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 7, pp. 3187–3197, Jul. 2023, doi: 10.1109/JBHI.2023.3264521.
- [36] W. Li, W. Huan, B. Hou, Y. Tian, Z. Zhang, and A. Song, "Can Emotion Be Transferred?—A Review on Transfer Learning for EEG-Based Emotion Recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 3, pp. 833–846, Sep. 2022, doi: 10.1109/TCDS.2021.3098842.
- [37] X. Zhang *et al.*, "Individual Similarity Guided Transfer Modeling for EEG-based Emotion Recognition," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA: IEEE, Nov. 2019, pp. 1156–1161. doi: 10.1109/BIBM47256.2019.8982972.
- [38] H. Cai and J. Pan, "Two-Phase Prototypical Contrastive Domain Generalization for Cross-Subject EEG-Based Emotion Recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096469.
- [39] Y.-P. Lin, "Constructing a Personalized Cross-Day EEG-Based Emotion-Classification Model Using Transfer Learning," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 5, pp. 1255–1264, May 2020, doi: 10.1109/JBHI.2019.2934172.
- [40] F. Wang *et al.*, "Emotion recognition with convolutional neural network and EEG-based EFDMs," *Neuropsychologia*, vol. 146, p. 107506, Sep. 2020, doi: 10.1016/j.neuropsychologia.2020.107506.
- [41] F. Liu *et al.*, "Emotion Recognition from Few-Channel EEG Signals by Integrating Deep Feature Aggregation and Transfer Learning," *IEEE Trans. Affective Comput.*, pp. 1–17, 2023, doi: 10.1109/TAFFC.2023.3336531.
- [42] D. Nizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China: IEEE, Jul. 2021, pp. 1–8. doi: 10.1109/IJCNN52387.2021.9534474.
- [43] R. Li, Y. Wang, W.-L. Zheng, and B.-L. Lu, "A Multi-view Spectral-Spatial-Temporal Masked Autoencoder for Decoding Emotions with Self-supervised Learning," in *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa Portugal: ACM, Oct. 2022, pp. 6–14. doi: 10.1145/3503161.3548243.
- [44] C. Rommel, J. Paillard, T. Moreau, and A. Gramfort, "Data augmentation for learning predictive models on EEG: a systematic comparison," *J. Neural Eng.*, vol. 19, no. 6, p. 066020, Dec. 2022, doi: 10.1088/1741-2552/aca220.
- [45] J. Shin, Y. Lee, S. Yoon, and K. Jung, "Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 2020, pp. 823–835. doi: 10.18653/v1/2020.acl-main.76.
- [46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*, vol. 119, no. 149, pp. 1597–1607, 2020.
- [47] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004, doi: 10.1016/j.jneumeth.2003.10.009.
- [48] X. Du *et al.*, "An Efficient LSTM Network for Emotion Recognition From Multichannel EEG Signals," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1528–1540, Jul. 2022, doi: 10.1109/TAFFC.2020.3013711.
- [49] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 29, 2017. Accessed: Dec. 30, 2023. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [50] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)." arXiv, Feb. 22, 2016. Accessed: Dec. 30, 2023. [Online]. Available: <http://arxiv.org/abs/1511.07289>
- [51] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A Multi-Modal Physiological Emotion Database for Discrete Emotion Recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019, doi: 10.1109/ACCESS.2019.2891579.
- [52] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From Regional to Global Brain: A Novel Hierarchical Spatial-Temporal Neural Network Model for EEG Emotion Recognition," *IEEE Trans. Affective Comput.*, vol. 13, no. 2, pp. 568–578, Apr. 2022, doi: 10.1109/TAFFC.2019.2922912.
- [53] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Systems with Applications*, vol. 47, pp. 35–41, Apr. 2016, doi: 10.1016/j.eswa.2015.10.049.
- [54] S. J. Reznik and J. J. B. Allen, "Frontal asymmetry as a mediator and moderator of emotion: An updated review," *Psychophysiology*, vol. 55, no. 1, p. e12965, Jan. 2018, doi: 10.1111/psyp.12965.
- [55] J. A. Coan and J. J. B. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biological Psychology*, vol. 67, no. 1–2, pp. 7–50, Oct. 2004, doi: 10.1016/j.biopsych.2004.03.002.

EmT: A Novel Transformer for Generalized Cross-subject EEG Emotion Recognition

Yi Ding, *Member, IEEE*, Chengxuan Tong, *Graduate Student Member, IEEE*, Shuailei Zhang, Muyun Jiang, Yong Li, Kevin Lim Jun Liang, and Cuntai Guan, *Fellow, IEEE*

arXiv:2406.18345v1 [cs.LG] 26 Jun 2024

Abstract—Integrating prior knowledge of neurophysiology into neural network architecture enhances the performance of emotion decoding. While numerous techniques emphasize learning spatial and short-term temporal patterns, there has been limited emphasis on capturing the vital long-term contextual information associated with emotional cognitive processes. In order to address this discrepancy, we introduce a novel transformer model called emotion transformer (EmT). EmT is designed to excel in both generalized cross-subject EEG emotion classification and regression tasks. In EmT, EEG signals are transformed into a temporal graph format, creating a sequence of EEG feature graphs using a temporal graph construction module (TGC). A novel residual multi-view pyramid GCN module (RMPG) is then proposed to learn dynamic graph representations for each EEG feature graph within the series, and the learned representations of each graph are fused into one token. Furthermore, we design a temporal contextual transformer module (TCT) with two types of token mixers to learn the temporal contextual information. Finally, the task-specific output module (TSO) generates the desired outputs. Experiments on four publicly available datasets show that EmT achieves higher results than the baseline methods for both EEG emotion classification and regression tasks. The code is available at <https://github.com/yi-ding-cs/EmT>

Index Terms—Deep learning, electroencephalography, graph neural networks, transformer.

I. INTRODUCTION

EMOTION recognition using electroencephalography (EEG) plays an important role in brain-computer interface (BCI) assisted mental disorder regulation. It requires machine to perceive human emotional states from brain activities using artificial intelligence techniques [1], [2]. The predictions of emotions can be used for neurofeedback in the regulation process [3]. Accurate predictions and fair generalization abilities to unseen subjects are crucial in building robust real-world BCI systems. Deep learning methods have shown promising results for accurate detection of brain activities [4], [5]. The design of the neural network architecture becomes crucial.

Incorporating neurophysiological prior knowledge into neural network architectures improves emotion decoding performance [6], [7], [8]. Common considerations include left-right hemisphere asymmetries, neurophysiologically meaningful graph connections, and the temporal dynamics of EEG

signals. According to neuropsychological studies, the left and right hemispheres react differently to emotions, particularly in the frontal areas [9]. Several deep learning methods [8], [10], [11] draw inspiration from this prior knowledge, achieving improved emotion decoding performance. EEG electrodes are placed on the scalp, which naturally forms a non-Euclidean structure. Therefore, many studies treat EEG signals as graphs, using either a neurophysiologically designed adjacency matrix [7] or a learnable one [12], [13]. However, there is still more prior knowledge that should be investigated. As one of the high-order cognitive processes in the brain, emotion consists of more basic processes such as attentional, perceptual, and mnemonic system processes [14]. Different brain regions are cooperatively activated under different cognitive processes, e.g., frontal and parietal networks in attention [15] and medial temporal lobes, prefrontal cortex, and parietal cortex interactions for episodic memory [16]. Using a pre-defined or a single learnable adjacency matrix cannot capture complex brain region connectives underlying emotional processes. Another prior knowledge about emotions is that emotional states are continuous in short periods while not consistent along the long stimuli [17]. Less attention is paid to this temporal contextual information underlying emotions.

To address the above-mentioned problems, we propose a novel transformer-based structure, named emotion transformer (EmT), for both generalized cross-subject emotion classification and regression tasks. To the best of our knowledge, this is the first work to explore transformer-based structures on cross-subject EEG emotion classification and regression tasks together. To learn temporal contextual information, we represent EEG segments as temporal graphs, as shown in Figure. 2. We learn the spatial information of each EEG graph to form a token. Then the long-short time contextual information is learned upon the token sequence.

We propose a residual multi-view pyramid graph convolutional neural networks (GCN) module, named RMPG, to capture multiple EEG channel connections for different basic cognitive processes in emotions. Multiple GCNs with independent learnable adjacency matrices are utilized parallelly in RMPG. Different from [18] that uses multiple adjacency matrices among manually defined local groups, we learn the connections among all the EEG channels. The reasons are two-fold: 1) locally defined groups are special cases of a global connection in which the EEG channels are fully connected within each local group, and 2) more specific channel-wise connections can be learned. Those parallel GCNs have different numbers of layers that can learn multi-view and

Yi Ding and Chengxuan Tong contribute equally to this work.

Yi Ding, Chengxuan Tong, Shuailei Zhang, Muyun Jiang, Yong Li, and Cuntai Guan are with the College of Computing and Data Science, Nanyang Technological University, 50 Nanyang Avenue, Singapore, 639798. e-mail: (ding.yi, tong0110, shuailei.zhang, james.jiang, yong.li, ctguan)@ntu.edu.sg.

Chengxuan Tong and Jun Liang Kevin Lim are with Wilmar International, Singapore. E-mail: (chengxuan.tong, kevin.limjunliang)@sg.wilmar-intl.com.

Cuntai Guan is the Corresponding Author.

multi-level graph embeddings with the help of their learnable adjacency matrices. Together with the output of a residual linear projection branch, a feature pyramid is formed. A mean fusion is utilized to combine the information in the feature pyramid as one token. Hence, a temporal sequence of tokens is formed.

We further propose temporal contextual transformers (TCT) with different token mixers [19] to learn contextual information from the token sequence for EEG emotion classification and regression tasks. An EEG trial refers to the period when one type of stimuli, e.g., a happy movie clip, is presented to the subject while the EEG is recorded. The label of the entire trial is assigned as happy according to either the self-assessments or the stimulus contents in classification tasks. However, the emotional states of the subjects change along with the stimuli [17]. Cutting each trial into short segments and assigning the same label to them will induce noisy labels [7]. To relieve noisy label issues and learn temporal contextual information of emotions, we can use a longer sliding window with short-shifting steps to split each trial, and the longer segments are further cut into sub-segments. Multi-head self-attention (MSA) in transformers [20] can attentively emphasize the parts that are highly correlated to the overall emotional state of the longer EEG segment. Because emotion is continuous [21], the underlying state is consistent in a short period. Hence, we propose a short-time aggregation (STA) layer after MSA to learn the long-short-time contextual information. Different from classification tasks, the label is temporally continuous in regression tasks. The model needs to regress the continuous changes of the emotional states for all the segments within a sequence. Although MSA can globally emphasize important parts in the sequence, recurrent neural networks (RNN) can further fuse the information of all the segments recurrently. This global information fusion ability makes RNN more suitable for regression task. Hence, we propose to use an RNN-based token mixer in TCT for the regression tasks instead of MSA. To this end, we propose TCT-Clas and TCT-Regr as the token mixers in EmT for EEG emotion classification and regression tasks.

The major contribution of this work can be summarised as:

- We propose a novel emotion transformer (EmT) for generalized cross-subject EEG emotion classification and regression tasks.
- We propose a residual multi-view pyramid GCN module (RMPG) to learn multi-view and multi-level graph embeddings incorporating the neuroscience knowledge that emotion consists of multiple basic cognitive processes.
- Two types of temporal contextual transformer (TCT) blocks with task-specific token mixers are proposed to learn temporal contextual information from EEG.
- Extensive experiments are conducted to evaluate and analyze the proposed EmT on four public datasets. Shanghai Jiao Tong University emotion EEG dataset (SEED) [22], the TsingHua University emotional profile dataset (THU-EP) [23] and the finer-grained affective computing EEG dataset (FACED) [24] are used for classification tasks while the multimodal database for affect recognition and implicit tagging dataset (MAHNOB-HCI) [25] is utilized

for regression tasks. The results demonstrate the superior of EmT over the compared baseline methods.

II. RELATED WORK

A. Graph Neural Networks

Graph neural networks (GNN) are used for non-Euclidean graph-structure data. Spectral GNN is a category of GNN that often rely on expensive eigendecomposition of the graph laplacian, thus several methods use approximation approaches to perform spectral filtering. ChebyNet [26] uses Chebyshev polynomials to approximate the spectral filters. Cayley polynomials are utilized to compute spectral filters for targeted frequency bands, which are localized in space and scale linearly with input data for sparse graphs [27]. The graph convolutional network (GCN) approximates spectral filtering with localized first-order aggregation [28]. EEG signals naturally have a graph structure. While some methods use GNN/GCN to extract spatial information from EEG signals, several approaches [7], [12], [13] do not consider the interactions among multiple brain areas involved in high-level cognitive processes, relying instead on a single adjacency matrix. Additionally, most of these methods [7], [12], [13], [18] neglect the temporal contextual information associated with emotional processes, using averaged features as node attributes. To effectively extract spatial relationships among EEG channels, we employ ChebyNet as the GCN layer in our model. Moreover, instead of solely using averaged features, we construct EEG signals as a sequence of spatial graphs to explicitly learn the temporal contextual information.

B. Temporal Context Learning

Emotion is a continuous cognitive process underlying which the temporal contextual information is embedded in the EEG signals. Methods that can learn the temporal dynamics of the sequence are often used as a temporal context extraction module. [25] utilize a long short-term memory network (LSTM) to predict the temporally continuous emotion scores. Using a temporal convolutional network (TCN) shows improved emotion regression results in [29]. However, both of them learn from flattened EEG feature vectors, which cannot effectively learn the spatial relations. TESANet [30] uses 1-D CNN to extract spatial information in spectral filtered EEG after which LSTM and self-attention are utilized to extract temporal dynamics to predict the odor pleasantness from the EEG signal. Conformer [31] combines CNN and transformer, achieving promising classification results for both emotion and motor imagery tasks. AMDET [32] utilizes a transformer and attention mechanism on the spectral-spatial-temporal dimensions of EEG data for EEG emotion recognition. Different from them, our model effectively learns from spatial topology information via parallel GCNs with learnable adjacency matrices. We also propose a short-time aggregation layer inspired by the prior knowledge that emotion is short-term stable and long-term varying.

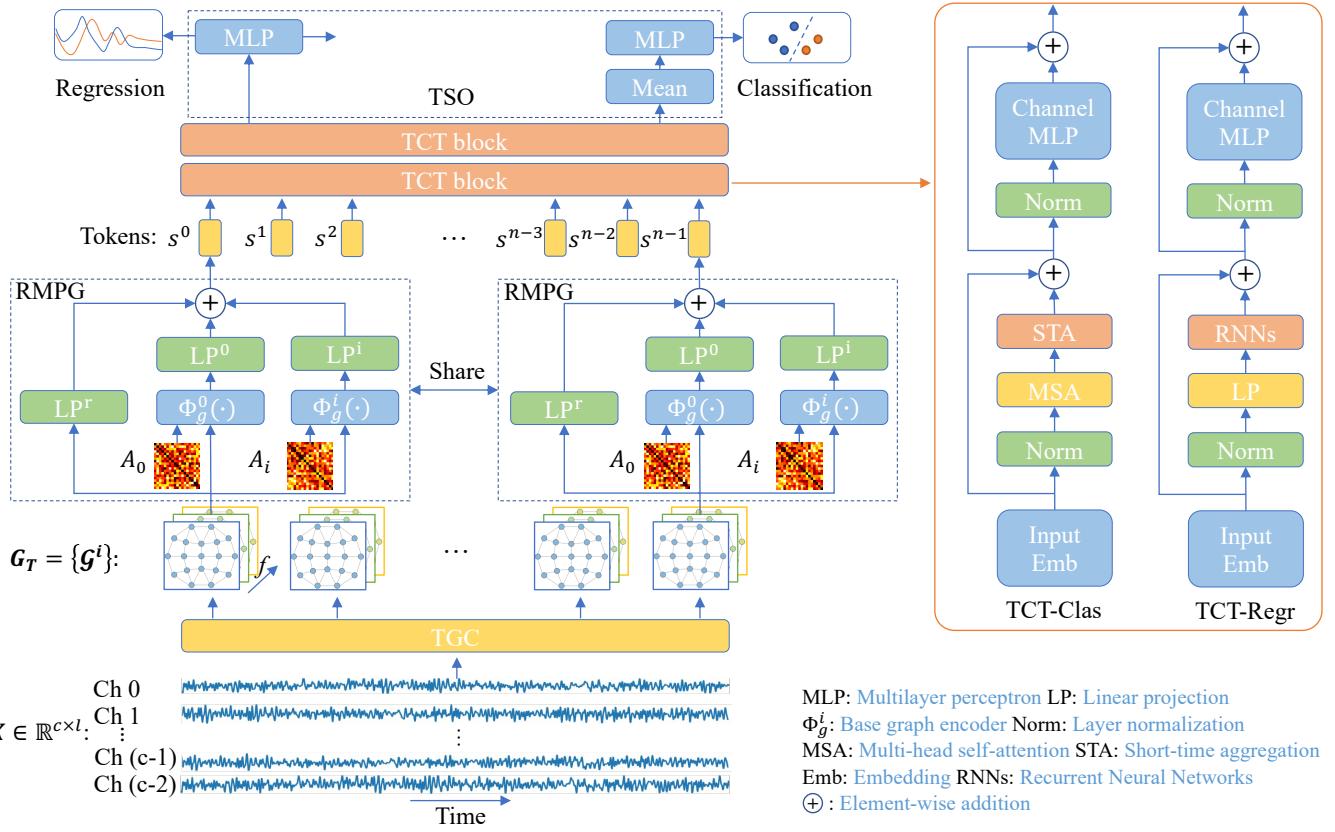


Fig. 1. The network structure of EmT. The temporal graphs from TGC are used as the input to RMPG that will transfer each graph into one token embedding. Then TCT extract the temporal contextual information via specially designed token mixers. We propose two types of TCT structures, named TCT-Clas and TCT-Regr, for classification and regression tasks separately. A mean fusion is applied before feeding the learned embeddings into MLP head for the classification output. For regression tasks, a MLP head projects each embedding in the sequence into a scalar to generate a sequence that can be used to regress the temporally continuous labels.

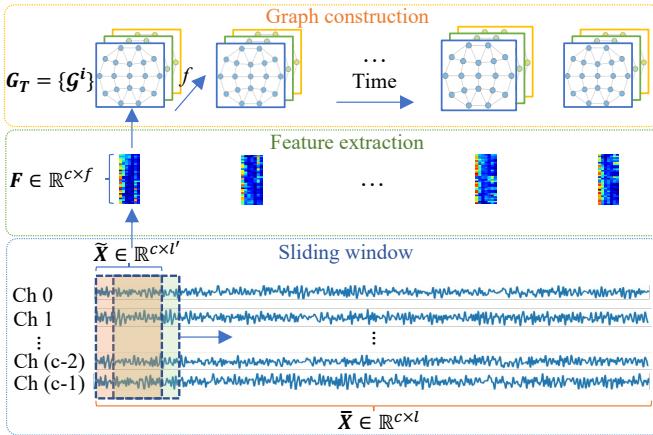


Fig. 2. Illustration of TGC. Each segment, \bar{X} , is split into several sub-segments, \tilde{X} . Features in different frequency bands are extracted for each \tilde{X} channel by channel to form F . Then each EEG channel is regarded as a node, and the extracted features are treated as node attributes. Combing all the graphs which are in time order, we get the temporal graphs, G_T .

C. EEG Emotion Recognition

Emotion recognition using EEG data presents a formidable challenge, primarily due to the inherent variability across subjects and the subjectivity involved in perceiving emotions.

The efficacy of prediction hinges on the development of a model capable of discerning crucial features that can distinguish between emotional classes. These extracted features typically encompass spectral, spatial, and temporal characteristics. Spectral features are typically derived through Fourier-based techniques, involving data filtering, and subsequent calculation of parameters like power spectral density (PSD) or differential entropy (DE). Alternatively, spectral filtering can be achieved by convolutions along the temporal dimension. Spatial features, on the other hand, are obtained through spatial convolution, often using convolutional neural networks (CNNs) and graph convolutional neural networks (GNNs). [12] propose DGCNN, a GCN-based network with a learnable adjacency matrix, to learn dynamical spatial patterns from differential entropy (DE) features. Based on DGCNN, a broad learning system (BLS) is added in graph convolutional broad network (GCB-Net) [13], which improves the emotion classification results. [7] proposed a regularized graph neural network (RGNN) with a neuroscience-inspired learnable adjacency matrix that is constrained to be symmetric and sparse to perform graph convolution. [18] proposed the instance-adaptive graph method to create graphic connections that are adapted to the given input. TSception [8] utilizes multi-scaled temporal and spatial kernels to extract multiple frequency and

spatial asymmetry patterns from EEG. Although they can learn the spatial and short-time temporal patterns, less attention is paid to the long-short-time contextual information underlying emotional cognitive processes.

III. METHOD

In this work, we propose a novel transformer, EmT, for both the generalized cross-subject EEG emotion classification and regression tasks. The network architecture is shown in Figure 1. EmT consists of four main parts: (1) temporal graph construction module (TGC), (2) RMPG, (3) TCT, and (4) task-specific output module (TSO). In TGC, EEG signals are transformed into temporal graph format that is a sequence of EEG feature graphs, as shown in Figure 2. RMPG learns dynamical graph representations for each EEG feature graph within the series and the learned representations of each graph are fused into one token. TCT learns the temporal contextual information via specially designed token mixers. Finally, the TSO module will generate the desired output for classification and regression tasks accordingly.

A. EEG-temporal-graph Representations

Temporal graphs are constructed to allow the neural network to learn spatial and temporal contextual information. Two steps are required to generate the temporal graphs from EEG, which are EEG segment/sub-segment segmentation and feature extraction, respectively. Figure 2 shows the construction process for one EEG segment.

Firstly, EEG signals are split into short segments that are further split into several shorter sub-segments using sliding windows. Given one trial of c -channel EEG signals, denoted by $\mathbf{X} \in \mathbb{R}^{c \times L}$, it is split into short segment $\tilde{\mathbf{X}} \in \mathbb{R}^{c \times l}$ using a sliding window of length l , with a hop step being s . Then another sliding window whose length and hop step are l' and s' is utilized to split each $\tilde{\mathbf{X}}$ into a series of sub-segment $\tilde{\tilde{\mathbf{X}}} \in \mathbb{R}^{c \times l'}$, where $L > l > l'$. In this paper, $l = 20\text{sec} = 20 * f_s$, $s = 4\text{sec} = 4 * f_s$ for all three datasets, where f_s is the sampling rate of EEG signals. For SEED and THU-EP, $l' = 2\text{sec} = 2 * f_s$ and $s' = 0.5\text{sec} = 0.5 * f_s$. For FACED, $l' = 4\text{sec} = 4 * f_s$ and $s' = 1\text{sec} = 1 * f_s$.

Relative power spectral density (rPSD) features are calculated for each sub-segment $\tilde{\tilde{\mathbf{X}}}$ to extract frequency information of short-period EEG signals. Specifically, the rPSD in delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), low beta (12-16 Hz), beta (16-20 Hz), high beta (20-28 Hz), and gamma (30-45 Hz) seven bands are calculated using Welch's method for each EEG channel to get a feature matrix $\mathbf{F} \in \mathbb{R}^{c \times f}$, where $f = 7$, for each $\tilde{\tilde{\mathbf{X}}}$. Each channel is regarded as one node and the rPSDs are regarded as node attributes. Hence, we have a temporal graph representation $\mathbf{G}_T = \{\mathcal{G}^i\} \in \mathbb{R}^{seq \times c \times f}$ for one $\tilde{\tilde{\mathbf{X}}}$.

B. Residual Multi-view Pyramid GCN

An RMPG is proposed to modulate the dynamical spatial relations among EEG channels underlying emotional processes. For each \mathcal{G}^i in \mathbf{G}_T , RMPG learns one flattened embedding as one temporal token \mathbf{s}^i for the subsequent transformer model.

A base graph encoder, $\Phi_g(\cdot)$, is utilized to learn graph representations. $\Phi_g(\cdot)$ can be ChebyNet [26], GCN [28], GAT [33] etc.. In this paper, we use ChebyNet as our base graph encoder:

$$\Phi_g(\mathbf{F}^m, \mathbf{A}) = \sigma \left(\sum_{k=0}^{K-1} \theta_k^m T_k(\hat{\mathbf{L}}) \mathbf{F}^{m-1} - \mathbf{b}^m \right), \quad (1)$$

where $m = [1, 2, \dots]$ is the number of GCN layers, $\mathbf{A} \in \mathbb{R}^{c \times c}$ is the adjacency matrix, σ is the ReLU activation function, θ is the learnable parameter, T_k is the k -order Chebyshev polynomials that are defined as $T_0(x) = 1$, $T_1(x) = x$, $T_i(x) = 2T_{i-1}(x) - T_{i-2}(x)$, $\hat{\mathbf{L}} = 2\mathbf{L}/\lambda_{max} - \mathbf{I}$ with $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$ is the re-scaling operation of \mathbf{L} , and \mathbf{b} is the bias. We approximate $\lambda_{max} \approx 2$ as [28] to remove its high computational cost. Hence, we have $\hat{\mathbf{L}} = -\mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$.

To modulate different brain region connections for multiple basic cognitive processes underlying emotions [14], we propose to use multiple different-layer GCNs, $\{\Phi_g^0(\cdot), \Phi_g^1(\cdot), \dots, \Phi_g^i(\cdot)\}$ with learnable adjacency matrices [12], $\{\mathbf{A}^0, \mathbf{A}^1, \dots, \mathbf{A}^i\}$. Each \mathbf{A}^i can learn one view of graph connections that belongs to a certain basic cognitive process. For each $\Phi_g^i(\cdot)$, stacking different layers of GCNs can learn different degrees of node cluster similarity [34]. Intuitively, for localized connections, such as electrodes within a brain functional area, a deeper GCN can get a consistent representation among these nodes. However, for global connections that are among different brain functional areas, a shallow GCN can aggregate the information among these areas while not over smooth it. The deeper the $\Phi_g^i(\cdot)$ is, the taller the feature pyramid level its output has. A linear projection layer, $\text{LP}(\cdot)$, is added for each $\Phi_g^i(\cdot)$ to project the flattened graph representations into a hidden embedding denoted by $\mathbf{H}_g^i \in \mathbb{R}^{d_g}$. By stacking \mathbf{H}_g^i from parallel different-layer GCNs, we can get multi-view pyramid graph embeddings as:

$$\{\mathbf{H}_g^i\} = \{\text{LP}^i(\Gamma(\Phi_g^i(\mathbf{F}, \mathbf{A}^i)))\}, \quad (2)$$

where $\{\cdot\}$ is the stack operation and $\Gamma(\cdot)$ is the flatten operation.

A linear residual branch is added to additionally provide some information about the non-filtered graphs. The output serves as the base of the feature pyramid. The linear layer project the flattened \mathcal{G}^i into a vector, \mathbf{H}_{g-base} , that has the same size as \mathbf{H}_g^i . A mean fusion is applied to combine different level graph information in the feature pyramid to form one token:

$$\mathbf{s} = \text{mean}(\{\mathbf{H}_{g-base}, \mathbf{H}_g^0, \dots, \mathbf{H}_g^i\}), \quad (3)$$

where $\mathbf{s} \in \mathbb{R}^{d_g}$ is the token embedding for each \mathcal{G} of \mathbf{G}_T , and $\mathbf{H}_{g-base} = \text{LP}(\Gamma(\mathcal{G}))$. To this end, \mathbf{G}_T becomes a temporal token sequence denoted by $\mathbf{S}_T = \{\mathbf{s}^i\} \in \mathbb{R}^{seq \times d_g}$.

C. Temporal Contextual Transformer

Two types of TCTs with different token mixers for classification and regression tasks are proposed to capture temporal contextual information of EEG underlying emotional processes. MetaFormer [19] shows that MSA can be replaced by different token mixers in transformers. We propose two

types of token mixers with which we have EmT-Clas and EmT-Regr for EEG emotion classification and regression separately. EmT refers to EmT-Clas unless otherwise stated. Given $S_T = \{s^i\} \in \mathbb{R}^{seq \times d_g}$ from RMPG, one block of TCT can be represented as:

$$\mathbf{Z}^m = \text{TokenMixer}_{\text{clas/regr}}(\text{Norm}(\mathbf{Z}^{m-1})) + \mathbf{Z}^{m-1}, \quad (4)$$

$$\mathbf{Z}^{m+1} = \text{MLP}(\text{Norm}(\mathbf{Z}^m)) + \mathbf{Z}^m, \quad (5)$$

where $m = [1, 2, \dots]$ is the number of layers in TCT blocks, $\mathbf{Z}^0 = S_T$, and MLP has two linear layers with the ReLU activation in between. A dropout layer is added after each linear layer.

1) *Token Mixers for Classification Tasks*: For classification tasks, MSA is utilized in $\text{TokenMixer}_{\text{clas}}(\cdot)$ to attentively emphasize the parts that are highly correlated to the overall emotional state of S_T . The tokens in S_T are then linearly projected into multiple groups of key (K^i), query (Q^i), and value (V^i) using multiple LP(\cdot), parameterized by $\mathbf{W}_{qkv}^i \in \mathbb{R}^{d_g \times 3d_{head}}$:

$$\{Q^i, K^i, V^i\} = \text{LP}^i(S_T) = S_T \mathbf{W}_{qkv}^i. \quad (6)$$

The scaled dot-product is utilized as the attention operation along temporal tokens to capture long-time context:

$$\text{Attn}(Q, K, V) = \text{Softmax}(QK^T / \sqrt{d})V, \quad (7)$$

where $d = d_{head}$ is a scaling factor. Because we need to apply the proposed STA on separate outputs from heads, we just stack the head outputs in the formula below:

$$\text{MSA}(S_T) = \{\text{Attn}(\text{LP}^0(S_T), \dots, \text{Attn}(\text{LP}^{n_{head}-1}(S_T)\}). \quad (8)$$

Considering the fact that emotion is short-term continuous and long-term varying [17], we propose a short-time aggregation (STA) layer after MSA to learn the long-short-time contextual information. A dropout layer with a scaling factor α , $\text{dp}(\alpha)$ is applied to control the forgetting rate of the temporal context. α is a hyper-parameter that will scale down the overall dropout rate in STA. To capture short-term consist patterns, CNN kernels denoted by \mathbf{K}_{cnn} whose size and step are $(n_{anchor}, 1)$, $(1, 1)$ are utilized to aggregate n_{anchor} temporal neighbors after MSA. Let $\mathbf{H}_{attn} \in \mathbb{R}^{n_{head} \times seq \times d_{head}}$ denotes one output of MSA. STA can be described as:

$$\text{STA}(\mathbf{H}_{attn}) = \text{Reshape}(\text{Conv2D}(\text{dp}(\mathbf{H}_{attn}), \mathbf{K}_{cnn})) \mathbf{W}_{sta}, \quad (9)$$

where Conv2D(\cdot) is the 2D convolution operation with the input being $\text{dp}(\mathbf{H}_{attn})$ and the kernel being \mathbf{K}_{cnn} , Reshape(\cdot) is the reshape operation $(n_{head}, seq, d_{head}) \rightarrow (seq, n_{head} * d_{head})$, and $\mathbf{W}_{sta} \in \mathbb{R}^{n_{head} * d_{head} \times d_g}$ is the projection weight matrix. The input and output channels of the Conv2D(\cdot) all equal n_{head} . The same padding is utilized to avoid size changes before and after STA. Hence, $\text{TokenMixer}_{\text{clas}}$ can be described as:

$$\text{TokenMixer}_{\text{clas}}(S_T) = \text{STA}(\text{MSA}(S_T)). \quad (10)$$

TABLE I
DETAILS OF EMT VARIANTS.

Model	GCN layers*	d_g	TCT layers	n_{head}	d_{head}	K^\dagger
EmT-S	1, 2	32	2	16	32	3/4
EmT-B	1, 2	32	4	16	32	3/4
EmT-D	1, 2	32	8	16	32	3/4

*: the layers of two parallel GCNs; \dagger : $K = 3$ for less-channel EEG (e.g. 32 in THU-EP and MAHNOB-HCI) and $K = 4$ for more-channel EEG (e.g. 62 in SEED).

2) *Token Mixers for Regression Tasks*: An RNN-based token mixer is utilized instead of MSA for $\text{TokenMixer}_{\text{regr}}(\cdot)$. MSA emphasises the parts that are highly correlated to the overall emotional state within a sequence via SA. This is helpful for the classification task that requires a single output of the sequence. However, the regression task needs the model to predict the continuous changes in the emotional states for all the segments within the sequence. Because RNN-based token mixer can fuse the information of all the segments recurrently, it is more suitable for the regression task. The tokens in $S_T \in \mathbb{R}^{seq \times d_g}$ are projected into values (V) using a projecting weight matrix, $\mathbf{W}_v \in \mathbb{R}^{d_g \times d_{head}}$. Because the RNN-based token mixer can be RNN, LSTM, GRU, etc., we use RNNs to denote the RNN family. A two-layer bi-directional GRU whose output length is $2 * d_{head}$ is empirically selected as the token mixer for EmT-Regr in this paper. Hence, $\text{TokenMixer}_{\text{regr}}$ can be described by:

$$\text{TokenMixer}_{\text{regr}}(S_T) = \text{RNNs}(\text{LP}(S_T)), \quad (11)$$

where $\text{LP}(S_T) = S_T \mathbf{W}_v$ is a linear layer.

D. Task-specific Output Module

MLP heads are utilized to generate the desired output for classification and regression tasks. Let $S_{\text{clas}} \in \mathbb{R}^{seq \times d_{head}}$ and $S_{\text{regr}} \in \mathbb{R}^{seq \times 2 * d_{head}}$ denote the learned embedding sequences for classification and regression tasks. The difference is that a mean fusion is applied to S_{clas} to combine the information of all the segments. Hence, the final classification output, $\hat{Y}_{\text{clas}} \in \mathbb{R}^{n_{class}}$ is calculated by:

$$\hat{Y}_{\text{clas}} = \text{mean}(S_{\text{clas}}) \mathbf{W}_{\text{clas}} + \mathbf{b}_{\text{clas}} \quad (12)$$

where $\mathbf{W}_{\text{clas}} \in \mathbb{R}^{d_{head} \times n_{class}}$, and $\mathbf{b}_{\text{clas}} \in \mathbb{R}^{n_{class}}$ are the weights and bias, respectively. And the final regression output, $\hat{Y}_{\text{regr}} \in \mathbb{R}^{seq}$ is calculated by:

$$\hat{Y}_{\text{regr}} = S_{\text{regr}} \mathbf{W}_{\text{regr}} + b_{\text{regr}} \quad (13)$$

where $\mathbf{W}_{\text{regr}} \in \mathbb{R}^{2 * d_{head} \times 1}$, and $b_{\text{regr}} \in \mathbb{R}^1$ are the weights and bias, respectively.

IV. EXPERIMENT

A. Datasets

For the classification task, we evaluate the performance of EmT with three emotion EEG datasets, which are SEED [22], THU-EP [23] and FACED [24]. For the regression task, we use a subset of the MAHNOB-HCI dataset [25].

TABLE II

GENERALIZED EMOTION CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE SEED, THU-EP AND FACED DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD AND THE NEXT BEST ARE MARKED USING UNDERLINES.

Method	SEED				THU-EP				FACED			
	ACC	std	F1	std	ACC	std	F1	std	ACC	std	F1	std
DGCNN	0.724	0.145	0.619	0.311	0.567	0.033	0.647	0.052	0.562	0.045	0.697	0.043
GCB-Net	0.684	0.172	0.517	0.357	0.554	0.043	0.620	0.088	0.565	0.052	0.685	0.053
RGNN	<u>0.790</u>	0.148	<u>0.802</u>	0.133	0.572	0.030	0.695	0.054	0.587	0.050	0.722	0.721
TSception	0.662	0.181	0.621	0.283	0.591	0.059	<u>0.707</u>	0.060	0.619	0.088	0.702	0.237
TCN	0.765	0.140	0.737	0.219	0.577	0.031	0.677	0.031	0.552	0.035	0.673	0.035
LSTM	0.733	0.158	0.670	0.274	0.558	0.035	0.626	0.062	0.568	0.063	0.700	0.064
TESANet	0.658	0.115	0.606	0.240	0.578	0.037	0.673	0.072	0.593	0.054	0.723	0.056
Conformer	0.612	0.127	0.529	0.221	0.601	0.038	0.691	0.042	0.590	0.035	0.720	0.035
AMDET	0.721	0.168	0.649	0.306	0.586	0.044	0.677	0.098	0.591	0.043	<u>0.726</u>	0.043
EmT-S (ours)	0.780	0.117	0.759	0.159	0.570	0.032	0.662	0.050	0.589	0.029	0.723	0.030
EmT-B (ours)	0.788	0.117	0.793	0.120	<u>0.595</u>	0.047	0.724	0.044	<u>0.608</u>	0.065	0.740	0.058
EmT-D (ours)	0.802	0.115	0.821	0.093	0.583	0.049	0.671	0.102	0.579	0.068	0.709	0.062

The SEED dataset encompasses data from 15 native Chinese subjects, with the objective of eliciting negative, positive, and neutral emotions using 15 Chinese film clips. Each film clip has a duration of approximately 4 minutes. Subsequently, the participants are tasked with providing self-evaluations regarding their emotional responses, considering dimensions such as valence and arousal after viewing these film clips. To record brain activity, EEG signals are acquired using a 62-channel electrode setup arranged in the 10-20 system, at a high sampling rate of 1000 Hz. The data undergo preprocessing, which includes applying a bandpass filter with a frequency range of 0.3 to 50 Hz.

The THU-EP dataset comprises data from 80 subjects, involving the use of 28 video clips as stimuli to elicit negative, positive, and neutral emotions. Each video clip has an average duration of approximately 67 seconds. These video clips are associated with a range of emotion items, including anger, disgust, fear, sadness, amusement, joy, inspiration, tenderness, arousal, valence, familiarity, and liking. After viewing each video clip, subjects provided self-report emotional scores for these emotion items. The EEG data was recorded using a 32-channel EEG system at a sampling rate of 250 Hz. The collected data underwent preprocessing, which included applying a bandpass filter with a frequency range of 0.05 to 47 Hz. Additionally, independent component analysis (ICA) was employed to effectively remove artifacts from the EEG data.

The FACED dataset is an extended version of the THU-EP dataset, comprising data from 123 subjects. The experimental protocol remains the same as THU-EP, with the addition of 43 subjects, making it a relatively larger dataset for studying emotions using EEG. The official pre-processed data is used in this study, with pre-processing steps identical to those in THU-EP.

MAHNOB-HCI is a comprehensive multi-modal dataset designed for the investigation of human emotional responses and the implicit tagging of emotions. This dataset involves the participation of 30 subjects in data collection experiments. In these experiments, each subject views 20 film clips while a variety of data streams are recorded in synchronization.

For the specific task of emotion recognition, a subset of the MAHNOB-HCI database, as referenced in [1], is employed. This subset encompasses 24 participants and 239 trials, with continuous valence labels provided by multiple experts. The final labels used for analysis are derived by averaging the annotations from these experts. The EEG signals in the dataset are collected using 32 electrodes, and they are sampled at a rate of 256 Hz. Notably, the annotations in the dataset are characterized by a resolution of 4 Hz.

B. Baselines

We demonstrate the performance of EmT by comparing the following baseline methods:

1) *DGCNN (graph-based)*: DGCNN [12] dynamically learns the relationships between different EEG channels through neural network training, represented by a learnable adjacency matrix. This dynamic learning enhances the extraction of more discriminative EEG features, ultimately improving EEG emotion recognition.

2) *GCB-Net (graph-based)*: Based on DGCNN, graph convolutional broad network (GCN-Net) [13] integrates a broad learning system (BLS) to the neural network. GCB-Net employs a graph convolutional layer to extract features from graph-structured input and then stacking multiple conventional convolutional layers to derive relatively abstract features. Subsequently, the final concatenation phase adopts a broad concept, preserving the outputs from all hierarchical layers, which facilitates the model in exploring features across a wide spectrum.

3) *RGNN (graph-based)*: RGNN [8] utilizes the biological topology of different brain regions to capture both local and global relationships among EEG channels. It models the inter-channel relations in EEG signals through an adjacency matrix within a graph neural network, where the connections and sparsity of the adjacency matrix are inspired by neuroscience theories regarding human brain organization.

4) *TSception (CNN-based)*: TSception [8] is a multi-scale convolutional neural network designed for generalized emotion recognition from EEG data. TSception incorporates dynamic temporal, asymmetric spatial, and high-level fusion layers,

working together to extract discriminative temporal dynamics and spatial asymmetry from both the time and channel dimensions.

5) *LSTM (temporal-learning)*: LSTM based neural networks [25] are utilized to predict the continuously annotated emotional labels. LSTM is capable to learn the long-term temporal dependencies among the EEG segments.

6) *TCN (temporal-learning)*: Zhang et al., use TCN [29] to learn the temporal information for the continuous EEG emotion regression tasks. The results indicate TCN has better regression performances than LSTM.

7) *TESANet (temporal-learning)*: TESANet [30] has been developed to discern the relations between time segments within EEG data, enabling the prediction of pleasant and unpleasant emotional states. This network architecture comprises a filter-bank layer, a spatial convolution layer, a temporal segmentation layer responsible for partitioning the data into overlapping time windows, a LSTM layer for encoding these temporal segments, and a self-attention layer.

8) *Conformer (temporal-learning)*: The Conformer [31] combines a CNN encoder and a transformer to capture both the short- and long-term temporal dynamics encoded in EEG signals, achieving promising results for emotion and motor imagery classification tasks.

9) *AMDET (temporal-learning)*: AMDET [32] learns from 3D temporal-spectral-spatial representations of EEG signals. It utilizes a spectral-spatial transformer encoder layer to extract meaningful features from the EEG signals and employs a temporal attention layer to highlight crucial time frames, considering the complementary nature of the spectral, spatial, and temporal features of the EEG data.

C. Experiment Settings

We conduct generalized subject-independent settings in this study where test data information is never used during the training stage. For SEED, we use a leave-one-subject-out (LOSO) setting. For each step in LOSO, one subject's data is selected as test data. 80% of the training data is used as training data, and the rest 20% is used as validation data. For the THU-EP and FACED datasets, we adopt a leave-n-subject-out setting [35], where $n_{\text{THU-EP}} = 8$, and $n_{\text{FACED}} = 12$. The last three subjects are added in the 10-th fold in FACED. 10% of the training data is used as the validation data. We perform binary classification on the positive and negative emotions on SEED, THU-EP, and FACED as [36]. We process the valence score of THU-EP and FACED into a binary class of high and low valence using a threshold of 3.0. For the regression task, we follow the same data pre-processing and experiment settings used in [29]. A LOSO is conducted for the regression task which is identical to the one for SEED.

D. Model Variants

We propose three types of EmT variants, namely EmT-shallow (EmT-S), EmT-base (EmT-B), and EmT-deep (EmT-D). The configurations are shown in Table I. As the names show, the differences among these variants are the depth of the TCT blocks. For the K of Chebyshev polynomials is

decided by the number of EEG channels (nodes in graphs). This is because K decides how many hops from the central vertex the GCN can aggregate. If the number of EEG channels is relatively less, e.g., 32 in THU-EP and MAHNOB-HCI, $K = 3$. K will be a larger value, $K = 4$ when the EEG signals have more channels such as 62 in SEED. EmT-S is used as the EmT-Regr because it gives the best results on the validation (development) set.

E. Evaluation Metrics

The evaluation metrics for emotion classification are the same as those in [8]: Accuracy (ACC) and F1 scores. They can be calculated by

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (14)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \quad (15)$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

The evaluation metrics for emotion regression are the same as those in [29]: root mean square error (RMSE), Pearson's correlation coefficient (PCC), and concordance correlation coefficient (CCC). Given the prediction $\hat{\mathbf{y}}$, and the continuous label \mathbf{y} , RMSE, PCC, and CCC can be calculated by

$$\text{RMSE} = \left\| \frac{\hat{\mathbf{y}} - \mathbf{y}}{N} \right\|^2 = \sqrt{\frac{\sum_{i=0}^{i=0} (\hat{y}_i - y_i)^2}{N}}, \quad (16)$$

$$\text{PCC} = \frac{\sigma_{\hat{\mathbf{y}}\mathbf{y}}}{\sigma_{\hat{\mathbf{y}}} \sigma_{\mathbf{y}}} = \frac{\sum_{i=0}^{i=0} (\hat{y}_i - \mu_{\hat{\mathbf{y}}})(y_i - \mu_{\mathbf{y}})}{\sqrt{\sum_{i=0}^{i=0} (\hat{y}_i - \mu_{\hat{\mathbf{y}}})^2} \sqrt{\sum_{i=0}^{i=0} (y_i - \mu_{\mathbf{y}})^2}}, \quad (17)$$

$$\text{CCC} = \frac{2\sigma_{\hat{\mathbf{y}}\mathbf{y}}}{\sigma_{\hat{\mathbf{y}}}^2 + \sigma_{\mathbf{y}}^2 + (\mu_{\hat{\mathbf{y}}} - \mu_{\mathbf{y}})}, \quad (18)$$

where N denotes the number of elements in the prediction/label vector, $\sigma_{\hat{\mathbf{y}}\mathbf{y}}$ denotes the covariance, $\sigma_{\hat{\mathbf{y}}}$ and $\sigma_{\mathbf{y}}$ are the variances, and $\mu_{\hat{\mathbf{y}}}$ and $\mu_{\mathbf{y}}$ are the means.

F. Implementation Details

The model configurations can be found in Table I. We first introduce the training parameters for classification tasks. The cross-entropy loss is utilized to guide the training. We use an AdamW optimizer with an initial learning rate of 3e-4. The label smoothing with a smoothing rate of 0.1 and a dropout rate of 0.25 are applied to avoid over-fitting. The batch size is 64 for all datasets. The training epochs are 10, 30, and 30 for SEED, THU-EP, and FACED. And the model with the best validation accuracy is used to evaluate the test data. α in STA is empirically selected as 0.25 for SEED, 0.1 for THU-EP and 0.4 for FACED. For regression tasks, we use a CCC loss, $\mathcal{L}_{\text{CCC}}(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \frac{2\sigma_{\hat{\mathbf{y}}\mathbf{y}}}{\sigma_{\hat{\mathbf{y}}}^2 + \sigma_{\mathbf{y}}^2 + (\mu_{\hat{\mathbf{y}}} - \mu_{\mathbf{y}})}$, where $\sigma_{\hat{\mathbf{y}}\mathbf{y}}$ is the covariance, $\sigma_{\hat{\mathbf{y}}}$ and $\sigma_{\mathbf{y}}$ are the variances, and $\mu_{\hat{\mathbf{y}}}$ and $\mu_{\mathbf{y}}$ are the means, an Adam optimizer with an initial learning rate of 5e-5 and a

TABLE III
EMOTION REGRESSION RESULTS ON MAHNOB-HCI.

Method	RMSE ↓	PCC ↑	CCC ↑
LSTM [25]	0.081	0.427	0.306
TCN [29]	0.066	0.474	0.377
EmT-Regr (MSA)*	0.075	0.393	0.312
EmT-Regr (LP+RNN)*	0.069	0.470	0.381
EmT-Regr (LP+LSTM)*	0.063	0.483	0.390
EmT-Regr (LP+GRU)*	0.068	0.490	0.396

↓: the lower the better; ↑: the higher the better.

*: EmT-Regr (token mixer type)

TABLE IV
GENERALIZED EMOTION CLASSIFICATION RESULTS OF ABLATION STUDIES ON THE SEED AND THU-EP DATASETS.

Method	SEED		THU-EP	
	ACC	F1	ACC	F1
w/o RMPG	0.775	0.793	0.577	0.663
w Single GCN	0.773	0.760	0.551	0.642
w/o TCT	0.777	0.749	0.592	0.698
w/o STA	0.784	0.795	0.582	0.690
EmT (ours)	0.802	0.821	0.595	0.724

weight decay of 1e-3, a batch size of 2, and a window length of 96 with a hop step of 32. We train the network for 30 epochs and use the model of the last epoch to evaluate the test data.

V. RESULTS AND ANALYSES

A. Emotion Classification

The experimental results are shown in Table II. We evaluate the methods using accuracy and F1 score. On the SEED dataset, EmT-D achieves the highest accuracy (0.802) and the highest F1 score (0.821), indicating its effectiveness in emotion classification. EmT-B also performs well with an accuracy of 0.788 and an F1 score of 0.793, while EmT-S shows a slightly lower accuracy of 0.780 but a strong F1 score of 0.759. RGNN achieves the second-best performance on the SEED dataset with an accuracy of 0.790 and an F1 score of 0.802. Notably, the methods that use features as input generally perform better than those using EEG as input. Additionally, learning from the temporal sequence of features consistently achieves better performance than learning from the features directly, except in the case of RGNN. This indicates the effectiveness of learning the temporal contextual information. Furthermore, compared to TCN, LSTM, TESANet, and AMDET, EmT can learn the spatial information better with the help of the GCN-based modules.

The observations on the THU-EP and FACED datasets are different. For the THU-EP dataset, EmT-B achieves the best F1 score of 0.724, while Conformer achieves the best accuracy (0.601). On the FACED dataset, EmT-B leads with the second-best accuracy (0.608) and the highest F1 score (0.740). AMDET achieves the second-best F1 score of 0.726, while the best accuracy is achieved by TSception (0.619). As the classes are imbalanced in both the THU-EP and FACED datasets, F1 scores are more important than accuracy. EmT-B achieves the

highest F1 scores on both THU-EP and FACED, demonstrating its effectiveness in emotion classification. Different from the observations on the SEED dataset, the baselines using EEG as input achieve better performance than those using features as input. This might be because there are more subjects in THU-EP and FACED, and using EEG directly can provide more information. Overall, the EmT series methods demonstrate the best overall performance across the evaluated datasets.

B. Emotion Regression

The regression results are shown in Table III. According to the results, EmT-Regr (LP+LSTM) achieved the lowest RMSE (0.063) among all compared approaches, while EmT-Regr (LP+GRU) achieved the best PCC (0.490) and CCC (0.396), indicating the effectiveness of the proposed method. However, when using MSA as the token mixer in EmT, the performance reduced dramatically, falling below all compared baselines. The difference between MSA and RNNs or TCN is that RNNs or TCN can fuse the information globally or locally, while MSA focuses on learning the global temporal relations and emphasizing certain parts of the sequence. Hence, the results indicate that fusing information from all segments is crucial for regression tasks.

C. Ablation

To explore the contributions of the RMPG, TCT, and STA modules, we conducted an ablation analysis by systematically removing each layer and observing the subsequent effects on classification performance. The results of this analysis are detailed in Table IV. Among all the ablation experiments, using only a single GCN had the most detrimental impact, leading to a 2.9% and 6.1% decrease in ACC and F1 score on the SEED dataset, and a 4.4% and 8.2% decrease on the THU-EP dataset. When EmT was tested without the RMPG module, it exhibited the second-largest reduction in accuracy on both datasets, with decreases of 2.7% and 1.8%, respectively. Similarly, the removal of TCT and STA modules also resulted in noticeable reductions in classification accuracy and F1 scores. These findings highlight that all the proposed modules work synergistically to enhance the predictive capabilities of EmT. Furthermore, the ablation results underscore the significant contribution of the RMPG module, attributable to its ability to modulate dynamic spatial relations among EEG channels, which is crucial for effectively capturing emotional activity.

D. Effect of EEG Features

Figure 3 illustrates the impact of different feature types on the SEED dataset for classification tasks and the MAHNOB-HCI dataset for regression tasks. We compared three types of features: PSD, DE, and rPSD, evaluating their effects on accuracy and F1 score for classification tasks, as well as on MRSE, CCC, and PCC for regression tasks. For classification tasks, rPSD outperformed the other two feature types, providing the highest accuracy and F1 score. Specifically, rPSD achieved 5.9% and 11.5% higher accuracy and F1 score, respectively, compared to EmT using DE. Furthermore, rPSD showed a

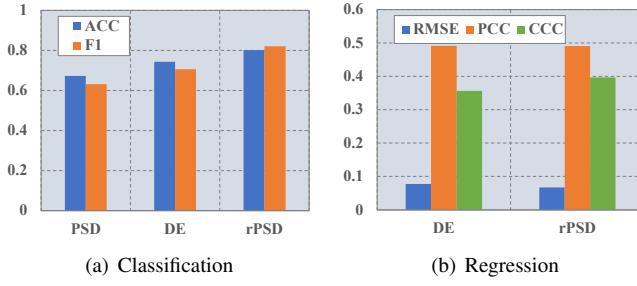


Fig. 3. Effect of feature types on emotion classification and regression performances of EmT using SEED and MAHNOB-HCI. Using rPSD gives the overall best performances. We don't add PSD results for regression tasks in (b) because the model cannot converge.

13.0% and 18.9% improvement in accuracy and F1 score, respectively, over PSD. These results suggest that rPSD is a superior spectral feature for EEG emotion classification tasks. For regression tasks, using rPSD resulted in better MRSE and CCC compared to using DE. However, the difference in PCC between rPSD and DE was minimal, with rPSD achieving a PCC of 0.491 and DE achieving a PCC of 0.490. Notably, when PSD was used as a feature, the model failed to converge, so we excluded these results from further analysis. Based on these findings, rPSD is demonstrated to be a more effective feature than DE, and both rPSD and DE are superior to PSD for EEG emotion recognition tasks.

E. Effect of The Number of TCT Blocks

We vary the number of TCT blocks from 2 to 8 and monitor their effect on the results of the two tasks. The results are shown in Figure 4. For classification, shown in Figure 4 (a), as the number of TCT blocks increases from 2 to 8, there is a noticeable improvement in both ACC and F1 scores. Specifically, accuracy increases from 0.780 to 0.802, while the F1 score shows a more substantial rise from 0.759 to 0.821, with the most significant improvement occurring between 6 and 8 TCT blocks. This indicates that adding more TCT blocks enhances the model's ability to capture temporal contextual information, thereby improving classification performance. Conversely, for regression tasks, shown in Figure 4 (b), the number of TCT blocks has little to no effect on the performance metrics. The RMSE remains stable around 0.06, and both the PCC and CCC show minimal variation, hovering around 0.48 and 0.39, respectively. This suggests that while increasing TCT blocks benefits classification by improving the capture of temporal contextual information, it does not significantly impact regression performance.

F. Visualization

Figure 5 illustrates the learned connectivity of brain regions on SEED. Two learnable adjacency matrices reveal different connectivity patterns during the emotional cognitive process. Figure 5 clearly shows a difference in the learned connectivities under emotion stimulation. In Figure 5 (a), the strongest learned connections are F6-POz, P2-Fz, F7-O1, PO5-Pz, T8-P7, and FC4-FPz. These connections indicate relationships

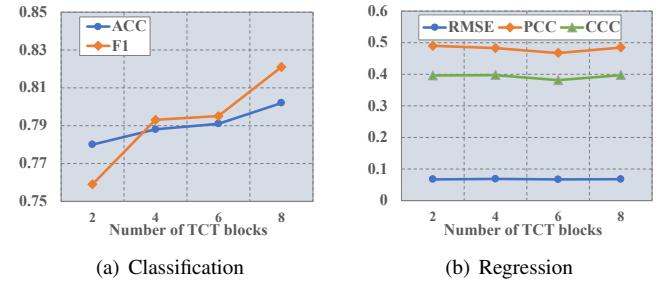


Fig. 4. Effect of the number of TCT blocks on emotion classification and regression performances of EmT using SEED and MAHNOB-HCI.

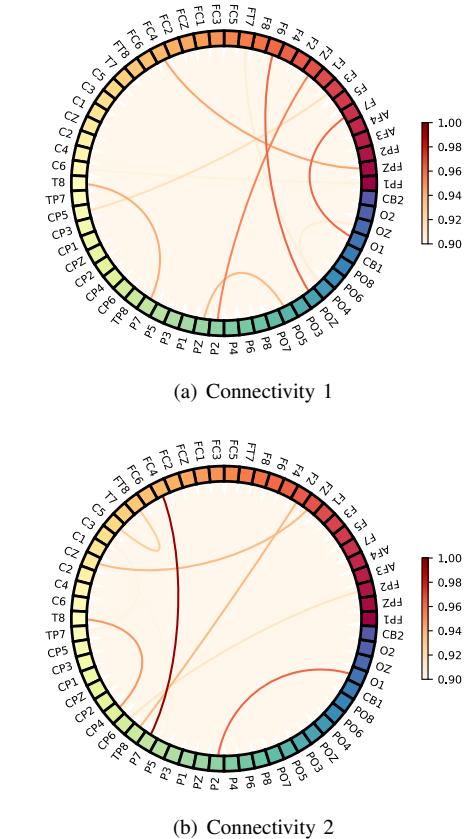


Fig. 5. Group-level connectivity of EEG channels for the emotional cognitive process. The two connectivity plots demonstrate that different connectivities of the brain are learned for the emotional task. (a) shows the first learned brain connectivity. (b) shows the second learned brain connectivity.

among the frontal, parietal, and temporal areas, which are known to be closely related to mental attention [37]. In Figure 5 (b), the important connections learned by EmT are FC4-P7, FT8-C5, Cz-Fz, T8-CP4, CP6-FP2, F2-TP8, and O1-P2. These connections include interactions among the frontal, temporal, and parietal areas, which are known to be related to emotions [38], [39], [40]. Additionally, some interactions involve the occipital and parietal areas, which are associated with visual processes. This is expected as videos were used as stimuli in the data collection experiments [22]. The two distinct patterns in Figure 5 reflect EmT's ability to capture multiple cognitive connectivities for the given classification task.

VI. CONCLUSION

In this paper, we propose a graph-transformer-based model, EmT, for cross-subject EEG emotion recognition. RMPG is proposed to learn multiple connection patterns for different brain cognitive processes under emotional stimulation. TCT is designed to learn temporal contextual information from the temporal EEG graphs. Subject-independent classification and regression tasks are conducted to evaluate EmT and relevant baseline methods. Results on three benchmarking datasets demonstrate EmT shows improvements over the compared baselines. Extensive analyses are performed to interpret the results and neurophysiologically meaningful features learned. Despite the improved emotion recognition performance, the learned temporal contextual information is difficult to identify. Proper experiments should be designed to pinpoint this temporal contextual information in the future. Additionally, in this paper, spatial connection information is learned using GCNs. Another useful tool for capturing connection patterns in EEG is the Riemannian manifold. Therefore, exploring how to capture contextual information using functional connection features and Riemannian manifolds would be an interesting direction for future research.

ACKNOWLEDGMENT

This work was supported by the RIE2020 AME Programmatic Fund, Singapore (No. A20G8b0102) and the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funding Scheme (project no. M23L8b0049) Scent Digitalization and Computation (SDC) Programme.

REFERENCES

- [1] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2019.
- [2] P. Li, H. Liu, Y. Si, C. Li, F. Li, X. Zhu, X. Huang, Y. Zeng, D. Yao, Y. Zhang, and P. Xu, "Eeg based emotion recognition by combining functional connectivity network and local activations," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2869–2881, 2019.
- [3] S. K. Ehrlich, K. R. Agres, C. Guan, and G. Cheng, "A closed-loop, music-based brain-computer interface for emotion mediation," *PloS one*, vol. 14, no. 3, p. e0213516, 2019.
- [4] J. C. Yudong Pan and Y. Zhang, "A survey of deep learning-based classification methods for steady-state visual evoked potentials," *Brain-Apparatus Communication: A Journal of Bacomics*, vol. 2, no. 1, p. 2181102, 2023.
- [5] E. Jeon, W. Ko, J. S. Yoon, and H.-I. Suk, "Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 2, pp. 739–749, 2023.
- [6] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong, "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 7 2018, pp. 1561–1567.
- [7] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2022.
- [8] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.
- [9] J. A. Coan and J. J. Allen, "Frontal eeg asymmetry as a moderator and mediator of emotion," *Biological psychology*, vol. 67, no. 1-2, pp. 7–50, 2004.
- [10] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong, "A novel neural network model based on cerebral hemispheric asymmetry for eeg emotion recognition," in *IJCAI*, 2018, pp. 1561–1567.
- [11] D. Huang, S. Chen, C. Liu, L. Zheng, Z. Tian, and D. Jiang, "Differences first in asymmetric brain: A bi-hemisphere discrepancy convolutional neural network for eeg emotion recognition," *Neurocomputing*, vol. 448, pp. 140–151, 2021.
- [12] 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, 2020.
- [13] T. Zhang, X. Wang, X. Xu, and C. L. P. Chen, "GCB-Net: Graph convolutional broad network and its application in emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 379–388, 2022.
- [14] H. Kober, L. F. Barrett, J. Joseph, E. Bliss-Moreau, K. Lindquist, and T. D. Wager, "Functional grouping and cortical–subcortical interactions in emotion: A meta-analysis of neuroimaging studies," *NeuroImage*, vol. 42, no. 2, pp. 998 – 1031, 2008.
- [15] P. Praamstra, L. Boutsen, and G. W. Humphreys, "Frontoparietal control of spatial attention and motor intention in human EEG," *Journal of Neurophysiology*, vol. 94, no. 1, pp. 764–774, 2005.
- [16] A. J. Watrous, N. Tandon, C. R. Conner, T. Pieters, and A. D. Ekstrom, "Frequency-specific network connectivity increases underlie accurate spatiotemporal memory retrieval," *Nature neuroscience*, vol. 16, no. 3, pp. 349–356, 2013.
- [17] Y. Zhang, H. Liu, D. Zhang, X. Chen, T. Qin, and Q. Zheng, "EEG-based emotion recognition with emotion localization via hierarchical self-attention," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.
- [18] T. Song, S. Liu, W. Zheng, Y. Zong, and Z. Cui, "Instance-adaptive graph for EEG emotion recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, pp. 2701–2708, Apr. 2020. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/5656>
- [19] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "MetaFormer is actually what you need for vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 819–10 829.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [21] P. Verduyn, P. Delaveau, J.-Y. Rotgé, P. Fossati, and I. V. Mechelen, "Determinants of emotion duration and underlying psychological and neural mechanisms," *Emotion Review*, vol. 7, no. 4, pp. 330–335, 2015.
- [22] 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.
- [23] 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, p. 118819, 2022.
- [24] J. Chen, X. Wang, C. Huang, X. Hu, X. Shen, and D. Zhang, "A large finer-grained affective computing eeg dataset," *Scientific Data*, vol. 10, p. 740, 2023. [Online]. Available: <https://doi.org/10.1038/s41597-023-02650-w>
- [25] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016.
- [26] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.
- [27] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, "CayleyNets: Graph Convolutional Neural Networks With Complex Rational Spectral Filters," *IEEE Transactions on Signal Processing*, vol. 67, no. 1, pp. 97–109, 2019.
- [28] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.
- [29] S. Zhang, C. Tang, and C. Guan, "Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition," *Pattern Recognition*, vol. 130, p. 108833, 2022.
- [30] C. Tong, Y. Ding, K. L. Jun Liang, Z. Zhang, H. Zhang, and C. Guan, "TESANet: Self-attention network for olfactory EEG classification," in

- 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–7.
- [31] Y. Song, Q. Zheng, B. Liu, and X. Gao, “EEG Conformer: Convolutional transformer for EEG decoding and visualization,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2023.
- [32] Y. Xu, Y. Du, L. Li, H. Lai, J. Zou, T. Zhou, L. Xiao, L. Liu, and P. Ma, “AMDET: Attention based multiple dimensions EEG transformer for emotion recognition,” *IEEE Transactions on Affective Computing*, pp. 1–11, 2023.
- [33] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in *International Conference on Learning Representations*, 2018.
- [34] Q. Li, Z. Han, and X.-m. Wu, “Deeper insights into graph convolutional networks for semi-supervised learning,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.
- [35] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, “Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2496–2511, 2023.
- [36] 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.
- [37] Y. Liu, J. Bengson, H. Huang, G. R. Mangun, and M. Ding, “Top-down Modulation of Neural Activity in Anticipatory Visual Attention: Control Mechanisms Revealed by Simultaneous EEG-fMRI,” *Cerebral Cortex*, vol. 26, no. 2, pp. 517–529, 09 2014.
- [38] Y. Gao, Z. Cao, J. Liu, and J. Zhang, “A novel dynamic brain network in arousal for brain states and emotion analysis,” *Mathematical Biosciences and Engineering*, vol. 18, no. 6, pp. 7440–7463, 2021.
- [39] D. Huang, C. Guan, K. K. Ang, H. Zhang, and Y. Pan, “Asymmetric spatial pattern for EEG-based emotion detection,” in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 1–7.
- [40] K. R. Mickley Steinmetz and E. A. Kensinger, “The effects of valence and arousal on the neural activity leading to subsequent memory,” *Psychophysiology*, vol. 46, no. 6, pp. 1190–1199, 2009.

MVGT: A Multi-view Graph Transformer Based on Spatial Relations for EEG Emotion Recognition

Yanjie Cui, Xiaohong Liu*, Jing Liang, Yamin Fu

School of Computer Science (National Pilot Software Engineering School)
Beijing University of Posts and Telecommunications, Beijing, 100876, China

{yanjiecui, xiaohongliu, liangjing18, fuyamin}@bupt.edu.cn

Abstract—Electroencephalography (EEG), a medical imaging technique that captures scalp electrical activity of brain structures via electrodes, has been widely used in affective computing. The spatial domain of EEG data is rich in affective information. However, few of the existing studies have simultaneously analyzed EEG signals from the perspectives of geometric and anatomical structures in spatial domain. In this paper, we propose a multi-view graph transformer (MVGT) based on spatial relations, which integrates information from the temporal, frequency and spatial domains, including geometric and anatomical structures, so as to enhance the expressive power of the model comprehensively. We encode the spatial information of EEG channels into the model, thereby improving its ability to comprehend the spatial structure of the channels. Experimental results from publicly available datasets demonstrate that our proposed model outperforms state-of-the-art baseline methods. Furthermore, the results also show that the MVGT could extract information from multiple domains and capture inter-channel relationships in EEG-based emotion recognition tasks effectively.

Index Terms—EEG, emotion recognition, graph transformer, structure encoding

I. INTRODUCTION

Affective computing is commonly employed for the analysis of emotional states through Human-Computer Interaction (HCI) systems, which collect multimodal data from subjects, including voice signals, self-reports, body gestures and physiological signals. Compared to other modalities, physiological signals have certain advantages. These signals are directly captured from the subjects' mental states, thus prevent subjects from disguising or hiding. The physiological signals commonly used to measure emotions are electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), and galvanic skin response (GSR), etc., among which EEG is often utilized to analyze the cognitive functions of human brain. Electrical signals from brain neurons are collected through dry and noninvasive electrodes placed on the scalp [1]. Nowadays, due to its high temporal resolution, portability, and affordability, this method is widely employed to study brain changes in response to emotional stimuli [2].

Traditional EEG features are mainly divided into three categories, i.e., time domain, frequency domain, and time-frequency domain features. Given the low signal-to-noise

ratio and substantial fluctuations inherent in EEG signals, frequency domain features are often used for EEG-based emotion recognition tasks. The typical approach involves decomposing the raw signals into five frequency bands: δ , θ , α , β , γ . Frequency domain features, such as power spectral density (PSD) [3], differential entropy (DE) [4], [5], differential asymmetry (DASM) [6] and rational asymmetry (RASM) [7], are subsequently extracted from each frequency band respectively.

The spatial structure of the brain also contains rich emotional information. Emotional states may involve distributed circuits rather than only a single brain region [8]. Asymmetry between the left and right hemispheres can reflect changes in valence and arousal [9]. Recent studies have highlighted the importance of utilizing spatial domain information. Li et al. [10] introduced recurrent neural networks to learn the asymmetric differences between the left and right hemispheres. Li et al. [11] also utilized hierarchical neural networks to learn both regional and global information of spatial-temporal EEG features. Graph neural networks (GNNs) are emerging as a powerful tool for analyzing spatial information in EEG emotion recognition. Song et al. [12] dynamically learned relationships between EEG channels using a graph convolutional network (GCN). Zhong et al. [13] incorporated asymmetry of the hemispheres into the adjacency matrix to model graph structure. Li et al. [14] also utilized an multi-domain adaptive graph convolutional network (MD-AGCN) to learn relationships between channels. Ding et al. [15] incorporated lobe information as prior knowledge into the GNN. Jiang et al. [16] proposed an elastic graph transformer (EmoGT) to extract emotional information. Although these methods have achieved excellent performance in emotion recognition, they have a common issue: they all rely on GNNs based on neighborhood aggregation schemes which may pose potential risks such as over-smoothing [17]–[19], under-reaching [20], and over-squashing [21]. Additionally, these methods do not take the geometric and anatomical structure information of the brain into account.

The main contributions of this paper lie in three aspects: (1) We propose a multi-view graph transformer (MVGT) based on spatial relations, fusing information from multiple perspectives including temporal, frequency, and spatial domains. (2) Our method, based on graph transformer, mitigates the

*Corresponding author

potential risks of over-smoothing, under-reaching and over-squashing occurring in traditional GNNs. Additionally, it enhances the model’s expressive power in emotion recognition by introducing spatial encodings based on geometric and brain lobe information. (3) Extensive experiments conducted on public datasets SEED and SEED-IV show our model achieves superior performance over the baseline models in emotion classification tasks.

II. RELATED WORK

In this section, we review the related work in terms of EEG-based emotion recognition and graph transformer.

A. EEG-based Emotion Recognition

EEG signals are inherently noisy and susceptible to channel crosstalk [22]. Due to the complexity of EEG signals, it is challenging to isolate clean and independent signals. Therefore, it is crucial to select what form of data to analyze under conditions of high noise. Effective features of EEG signals can reduce noise and facilitate the recognition of cognitive patterns in specific tasks. Experimental evidence suggests that frequency domain features are often associated with behavioral patterns [23], hence they are commonly used in EEG analysis.

Along with the development of deep learning, increasingly complex models with rich expressive abilities have emerged and have been extensively utilized in EEG signal analysis. Zheng et al. [5] employed a deep belief network to analyze important frequency domain components and effective channels based on the learned parameters. Song et al. [12] used a GCN method based on Chebyshev polynomials [24] to dynamically learn the representations of EEG signals. Zhong et al. [13] innovatively incorporated the inter-channel asymmetry of the hemispheres as prior knowledge into the adjacency matrix in 3D space. The reasonable combination of multi-domain information improves the accuracy in the emotion recognition tasks. Li et al. [14] proposed the MD-AGCN that integrates the temporal domain, frequency domain, and functional connectivity. Ding et al. [15], inspired by neuroscience research, combined intra-region convolution and inter-region convolution based on brain lobe information to learn cognitive patterns. Jiang et al. [16] utilized the advantages of GCN in the spatial domain and Transformer in the temporal domain to improve the accuracy of emotion classification.

B. Graph Transformer

The GNNs used in the methods above are based on neighborhood aggregation schemes. However, classical GNNs based on message passing (MPGNNs) may lead to over-smoothing [17]–[19], under-reaching [20], and may also fail to fit long-range signals due to over-squashing [21], which limit the expressive power of the model. Graph transformers (GTs) alleviate such effects as they have a global receptive field [25]. Nevertheless, without sufficiently expressive structural and positional encodings, GTs cannot capture effective graph structures [26]. Dwivedi et al. [27] utilized eigenvectors of graph Laplacian as position encodings in a fully connected

graph transformer and integrated edge features into the attention mechanism. Building on this, SAN [28] used a full Laplacian spectrum to learn the positional encodings for each node. Graphomer [29], [30] employed node centrality and node distance metric to implement structural and relative positional encodings, achieving state-of-the-art performance on challenging graph datasets. For EEG emotion recognition, Li et al. [31] innovatively combined a masked autoencoder based on self-supervised learning with a CNN-Transformer hybrid structure, effectively improving classification accuracy. However, this method only used the sine-cosine positional encoding, limiting the Transformer’s ability to learn spatial information.

III. PRELIMINARY

A. Graph Neural Network (GNN)

Let $G = (V, E)$ define a graph, where $V = \{v_1, v_2, \dots, v_n\}$ represents the nodes in the graph, and $E = \{e_1, e_2, \dots, e_m\}$ is the edges between the nodes. The representation of node v_i is denoted as $x_i \in \mathbb{R}^d$. Most existing GNNs [17], [32]–[35] adopt neighborhood aggregation schemes, iteratively aggregating representations of its first or higher-order neighbors, followed by using backpropagation (BP) to learn data-driven feature representations. We define the representation of node v_i at the l -th iteration as h_i^l and define $h_i^0 = x_i$. The l -th iteration can be represented as:

$$a_i^l = \text{AGGREGATE}^l (\{\varphi_\theta(h_j^{l-1}, e_{ji}) : j \in \mathcal{N}(v_i)\}), \quad (1)$$

$$h_i^l = \text{UPDATE}^l (h_i^{l-1}, a_i^l), \quad (2)$$

where φ_θ represents a differentiable function used for feature transformation of node and edge information. The set $\mathcal{N}(v_i)$ is the neighbors of v_i . The AGGREGATE function is used to aggregate the transformed representations using a differentiable, permutation invariant function (such as mean, sum, max, etc.). The goal of UPDATE function is to integrate the information from neighbors into the node representation. For graph classification, the READOUT operation is typically used to obtain a representation of the entire graph, which is then fed into a classifier to determine the graph label.

B. Graphomer

The Transformer [36] is undeniably one of the most popular deep neural network architectures today, driving significant advancements in natural language processing and computer vision. From the perspective of GNNs, Transformer can be interpreted as a GNN acting on a fully connected graph. Therefore, it is feasible to use the Transformer to address tasks on graph data. The ability to properly incorporate the structural information of graphs into the model is the key for leveraging its expressive power. Graphomer [29], [30] can go beyond classical MPGNNs in expressive power and achieves state-of-the-art performance on large molecular benchmarks. Graphomer incorporates centrality encoding into the graph data and integrates spatial encoding, edge encoding into the attention mechanism, which can be expressed as:

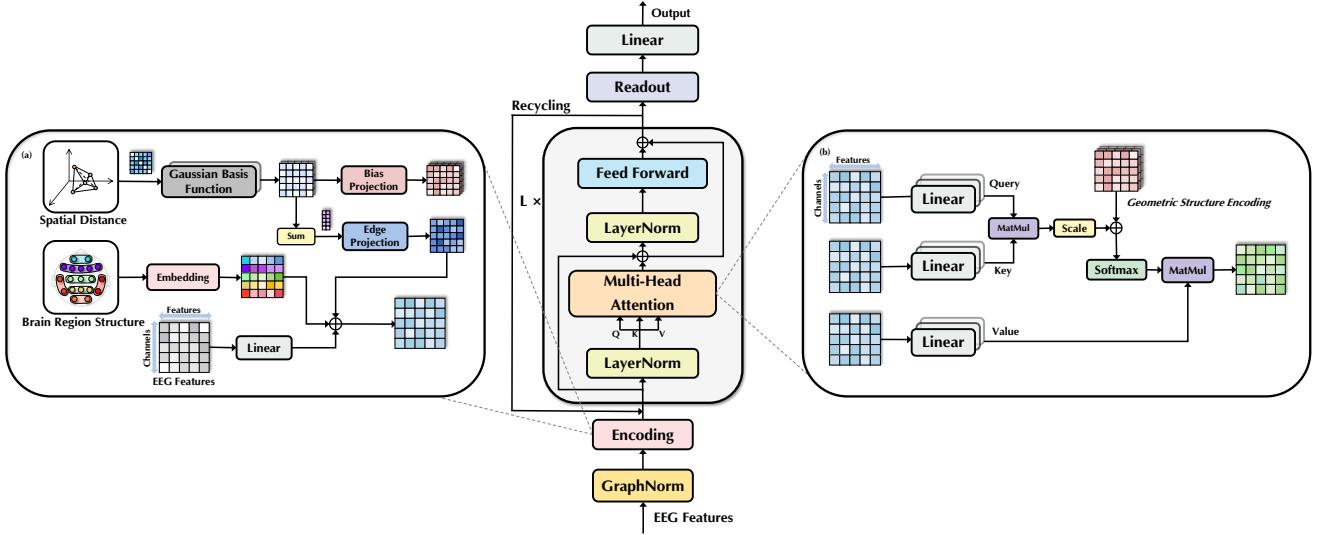


Figure 1. Overall structure of MVGT. (a) represents the encoding process of brain region structure and geometric structure. (b) depicts the process of calculating inter-channel correlations based on the attention mechanism and geometric structure encoding.

$$A_{ij} = \frac{(\mathbf{h}_i \mathbf{W}_Q)(\mathbf{h}_j \mathbf{W}_K)^\top}{\sqrt{d}} + b_{\phi(v_i, v_j)} + c_{ij}, \quad (3)$$

where c_{ij} represents the edge encoding on the shortest path and the bias term $b_{\phi(v_i, v_j)}$ can adaptively adjust the correlations between v_i and v_j .

IV. METHODS

In this section, we introduce the methods employed in the EEG emotion recognition task. First, we elaborate on the embedding of temporal information. Second, leveraging the spatial geometry and physiological anatomy of the brain, we propose two novel and simple designs of encoding that enable the model to adaptively learn the inter-channel correlations. Finally, we present the detailed implementations of MVGT.

A. Problem Definition

EEG signals can be represented as a two-dimensional matrix with respect to channels and time. Given that channels exhibit spatial structure, they can be structured into a fully connected graph $G = (V, E)$, where V denotes the nodes, representing EEG channels, and E denotes the edges, representing the connections between channels. The features of the nodes at time t are denoted by $\mathbf{X}_t = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \dots, \mathbf{x}_n^\top]^\top \in \mathbb{R}^{n \times d}$, where $n = |V|$ represents the number of nodes and d represents the feature dimension.

B. Temporal Embedding

EEG signals have high temporal resolution and contain rich temporal information. Because of the multi-electrode acquisition method, EEG signals can be regarded as multivariate time series. When processing time series, the embedding of temporal information are crucial. EmoGT treated the features of different channels at the same time points as tokens

and employed an attention mechanism to extract temporal correlations between them. Due to the anisotropic volume conduction characteristics [37] in human brain tissues, there may be temporal delays between different channels, which in turn leads to time-unaligned events at a single moment thus causing performance degradation. MD-AGCN utilized the convolutional operation to extract temporal information along the time axis from continuous EEG segments, with the receptive field limited by the size of the convolution kernel. Inspired by [38], we broaden the receptive field by considering the entire time series as an embedded token rather than a single time point. First, following the methods of MD-AGCN and EmoGT, we use overlapping sliding windows of size T to split EEG signals along the time axis and use these segments as samples, which are then fed into the attention module in the form of continuous segments. After processing with sliding windows, we obtain $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \dots, \tilde{\mathbf{X}}_S)$, where the s -th sample is $\tilde{\mathbf{X}}_s \in \mathbb{R}^{n \times Tf}$, S is the number of continuous EEG segments, n is the number of channels, and f is the dimension of frequency domain features.

According to the universal approximation theorem [39], the feed-forward neural network (FFN), as the basic module of the Transformer encoder, can learn the intrinsic properties to describe a time series and is a superior predictive representation learner compared to self-attention [38]. Therefore, using continuous time segments as the input to the FFN may be more effective in extracting the temporal information of each channel independently.

C. Spatial Encoding

The special structure of the brain encompasses rich spatial information. Fully exploiting structural information is beneficial for the recognition and analysis of cognitive patterns in the brain. Therefore, to better recognize emotional patterns

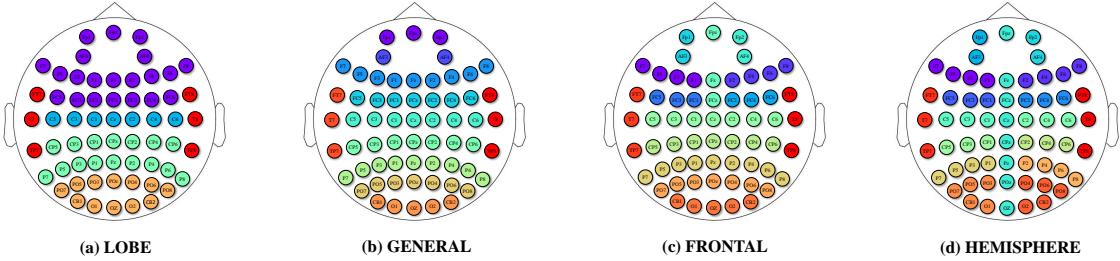


Figure 2. The brain region division scheme is illustrated. (a) LOBE scheme shows a coarse partitioning based on lobe structures. (b) GENERAL scheme represents a fine-grained partitioning of the brain lobes. (c) FRONTAL scheme introduces symmetry of the left and right frontal regions based on the GENERAL scheme. (d) HEMISPHERE scheme further enhances the channel symmetry in the partitioning scheme. Channels of the same color belong to the same brain region.

in emotion classification tasks, we employ two simple but effective methods of spatial encoding: brain region encoding and geometric structure encoding.

1) *Brain Region Encoding*: Neuroscience research demonstrated that the activation of a specific brain region often leads to the concurrent activation of related brain regions responsible for the same high-level cognition [40]. In EEG emotion recognition, incorporating relevant neuroscience findings can typically enhance recognition accuracy. Zhong et al. [13] integrated the asymmetry of neural activity between the left and right hemispheres as prior knowledge into the adjacency matrix, effectively enhancing recognition accuracy. Li et al. [10] improved emotion pattern recognition performance by capturing the differences between the left and right hemispheres. Ding et al. [15] divided EEG channels into different regions and combined local intra-region convolution with global inter-region convolution, achieving good results on the DEAP [41] dataset. With reference to the three divisions of [15], we adopt four brain region divisions, which divide the EEG channels into different regions based on a prior knowledge, aiming to incorporate the brain region information into the model. The division schemes are as follows:

- We divide the regions based on the anatomical structure of the brain and implement LOBE scheme.
- To further investigate the expressive power of brain region encoding, we conduct a detailed division of brain lobes according to the 10-20 system based on electrode positions, employing the GENERAL scheme.
- Asymmetric EEG activity in the frontal lobe can be utilized for discriminating valence changes [9]. The left frontal lobe exhibits a stronger correlation with joy and happiness, while the right frontal lobe is more strongly correlated with fear and sadness. Thus we further divide the frontal lobe region into two symmetrical regions to obtain the FRONTAL scheme.
- According to the symmetry of brain structure [42], we make a finer division of the brain lobe regions, defining the HEMISPHERE scheme.

The four schemes mentioned above are shown in Fig. 2. In terms of specific implementation, we assign a brain region tag to each electrode, then map the tags into an embedding space using a learnable projection function, and simply add

the embeddings to the node features. The encoding of node i is represented as follows:

$$\mathbf{r}_i = \text{Embedding}(\text{Tag}(\mathbf{x}_i)), \quad \mathbf{r}_i \in \mathbb{R}^d, \quad (4)$$

$$\mathbf{h}_i^0 = \mathbf{x}_i \mathbf{W}_{\mathcal{X}} + \mathbf{r}_i, \quad (5)$$

where $\mathbf{W}_{\mathcal{X}} \in \mathbb{R}^{Tf \times d}$ is a learnable projection function, and d represents the dimension of the embedding. Through the above encoding method, we integrate the brain's anatomy information into the model.

2) *Geometric Structure Encoding*: In the real world, the human reasoning process considers not only the semantic relationships between objects but also their spatial relations. EEG channels have a 3D structure, and the functional connectivity between these channels lacks precise definitions. Therefore, we represent the relationships between EEG channels as a fully connected directed graph. Firstly, let $\phi(i, j)$ represent the Euclidean distance between node i and node j , and encode $\phi(i, j)$ using a set of Gaussian basis functions [30], [43]. Let $\mathbf{b}_k \in \mathbb{R}^{n \times n}$ denote one of the Gaussian basis functions. The element (i, j) of this function can be expressed as:

$$\mathbf{b}_k(i, j) = \mathcal{G}_k(\alpha_{ij}\phi(i, j) + \beta_{ij} - \mu_k, \sigma_k), \quad (6)$$

where α_{ij} , β_{ij} , μ_k , and σ_k are learnable parameters, and i and j denote the index of the source and target node, respectively. The result of the basis functions can be represented as $\mathbf{B} = \|\mathbf{b}_1 \mathbf{b}_2 \dots \mathbf{b}_K\|$, with $\mathbf{B} \in \mathbb{R}^{n \times n \times K}$, where $\|\cdot\|$ denotes the concatenation. All geometric encodings of each node are then summed up along the second dimension and then transformed linearly.

$$\mathbf{h}_i^0 = \mathbf{x}_i \mathbf{W}_{\mathcal{X}} + \mathbf{z}_i \mathbf{W}_{\mathcal{Z}} + \mathbf{r}_i, \quad \mathbf{z}_{i,k} = \sum_{j=1}^n \mathbf{B}_{i,j,k}, \quad (7)$$

$$\mathbf{B}' = \text{Projection}(\mathbf{B}), \quad (8)$$

where i is the node index, k is the index of basis function and $\mathbf{W}_{\mathcal{Z}} \in \mathbb{R}^{K \times d}$ is a learnable projection matrix. Projection : $\mathbb{R}^{n \times n \times K} \mapsto \mathbb{R}^{n \times n \times M}$ is a nonlinear transformation, where M is the number of attention heads. We incorporate this encoding as a bias term into the softmax attention.

Our proposed spatial encoding matrix is directed, which is inconsistent with the assumption of a symmetric adjacency matrix [13], [16]. Using directed connections provides the model with greater expressive power because the correlation

between node pairs (i, j) and (j, i) may differ. Since we assume nodes are fully connected, we avoid specific assumptions about inter-channel correlations and learn the functional correlations between nodes through the encodings. Letting l denote the model depth, and i denote the index of multi-head attention, the functional brain connectivity of the s -th sample can be represented as:

$$\mathbf{A}_s^{l,i} = \text{Softmax} \left(\frac{\left(\mathbf{H}_s^l \mathbf{W}_Q^{l,i} \right) \left(\mathbf{H}_s^l \mathbf{W}_K^{l,i} \right)^T}{\sqrt{d_h^{l,i}}} + \mathbf{B}'^i \right), \quad (9)$$

where the projections are learnable parameters $\mathbf{W}_Q^{l,i} \in \mathbb{R}^{d \times d_h}$ and $\mathbf{W}_K^{l,i} \in \mathbb{R}^{d \times d_h}$. The scalar $d_h^{l,i}$ is the second dimension of $\mathbf{W}_K^{l,i}$. This encoding method integrates temporal, frequency, and spatial domain features into the model, enhancing its expressive power. We compute the attention scores between nodes using embedded vectors, representing the semantic correlations between different nodes from multiple perspectives. Finally, the attention scores are added to the spatial geometric encoding to obtain the correlations between channels.

D. Implementation Details of MVGT

In this section, we describe the overall architecture of the model, including spatial encodings and the Transformer encoder, as illustrated in Fig. 1. For better optimization, we first apply GraphNorm [44] to normalize the input features to a range between 0 and 1. Subsequently, we perform geometric and regional structure encodings to obtain multi-domain embeddings. The encodings could be characterized as below:

$$\mathbf{X}'_s = \text{GraphNorm}(\mathbf{X}_s) \quad (10)$$

$$\mathbf{H}_s^0 = \text{SpatialEncoding} + \text{Proj}(\mathbf{X}'_s) \quad (11)$$

We employ a Pre-LN Transformer structure, applying layer normalization (LN) before the multi-head attention (MHA) and the FFN. A recent study suggests that the Pre-LN structure yields more stable gradients and is more favorable for optimizer, enabling faster convergence [45] compared to Post-LN. Additionally, we utilize dropout to mitigate overfitting. This process is represented as follows:

$$\mathbf{H}_s^l = \text{MHA}(\text{LN}(\mathbf{H}_s^{l-1})) + \mathbf{H}_s^{l-1} \quad (12)$$

$$\mathbf{H}_s^l = \text{FFN}(\text{LN}(\mathbf{H}_s^l)) + \mathbf{H}_s^l \quad (13)$$

Inspired by [30], [46], we feed the outputs recursively into the same modules, denoted as "recycling" in Fig. 1. The iterative refinement progressively refines the model's ability to discriminate encoded information and understand emotional patterns, thereby helping the model capture more effective details.

V. EXPERIMENTS

A. Datasets

For our experiments, we select the SEED [5] and SEED-IV [47] datasets to evaluate the effectiveness of our model.

These datasets consist of EEG signals recorded from subjects watching emotion-eliciting videos.

SEED dataset comprises data from 15 subjects who participated in three sessions, each separated by at least one week. Each session consists of 15 trials capturing emotional labels, with the emotion labels being positive, negative, and neutral.

SEED-IV dataset is constituted by EEG signals from 15 subjects across three separate sessions conducted at different times, using the same device as the SEED dataset. This dataset encompasses four emotion labels: neutral, sad, fear, and happy. In each session, each subject underwent 24 trials.

B. Settings

To prevent potential data leakage that could arise from segment-wise shuffling, we split the training and test sets at the trial level. Following the settings of previous studies [5], [10], [12]–[14], [16], [31], [47], we use pre-computed differential entropy (DE) features for the recognition task. For the SEED dataset, we use the first 9 trials of each subject as the training set and the last 6 trials as the test set, as done in previous research. The DE features are computed using five frequency bands extracted from 1s nonoverlapping windows. The model performance is evaluated based on the average accuracy and standard deviation across all subjects over two sessions of EEG data. Similarly, for the SEED-IV dataset, we use the first 16 trials as the training set and the last 8 trials as the test set. The DE features for SEED-IV are calculated using 4s windows. The performance of our model is assessed using data from all

Table I
THE CLASSIFICATION ACCURACIES (MEAN/STD) ON SEED AND SEED-IV.
MVGT-L, MVGT-G, MVGT-H, MVGT-F: MVGT USING LOBE,
GENERAL, HEMISPHERE AND FRONTAL SCHEMES.

Model	SEED	SEED-IV
DGCNN [12]	90.40/08.49	69.88/16.29
BiHDM [10]	93.12/06.06	74.35/14.09
R2G-STNN [11]	93.34/05.96	-
RGNN [13]	94.24/05.95	79.37/10.54
MD-AGCN [14]	94.81/04.52	87.63/05.77
EmoGT [16]	95.02/05.99	91.20/09.60
MV-SSTMA [31]	95.32/3.05	92.82/5.03
MVGT-L	95.36/05.37	91.51/09.03
MVGT-G	94.43/05.35	93.57/08.60
MVGT-H	95.19/05.48	90.19/10.42
MVGT-F	96.45/04.40	91.62/09.05

three sessions.

For the input data, we use overlapping sliding windows of size T along the time axis to extract sample fragments, with T being set to 5. During experiments, the hidden dimension is set to 64 and the number of Gaussian basis functions is 32. The number of MHA layers is 4 and the number of attention heads is 2. The iterative refinement process is performed three times. We set the batch size to 32 and the learning rate within the range of 3e-5 to 3e-3. Cross-entropy is used as the loss function, and AdamW [48] is employed as the optimizer with a weight decay rate of 0.1.

C. Results Analysis

We compare the classification results based on the SEED and SEED-IV datasets with recent state-of-the-art models, as shown

in Table I. It is evident that our proposed model significantly outperforms the baseline models under the same experimental settings. For the SEED dataset, the model adopting the FRONTAL scheme achieves the best performance, with a classification accuracy of 96.45%. The LOBE scheme also achieves a slightly superior accuracy of 95.36%, compared to other models. For the SEED-IV dataset, the classification accuracy under the GENERAL scheme is 93.57%, achieving the best performance compared to baseline models. The MVGT model also demonstrates strong performance under other division schemes. Overall, our model achieves the best classification accuracy compared to the baselines. The results also suggest that selecting the specific division scheme relevant to the emotion task could enhance the expressive power of MVGT.

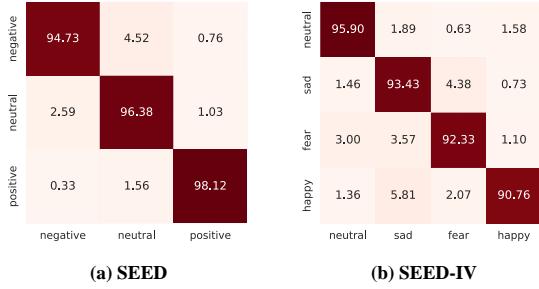


Figure 3. Confusion matrices of MVGT. (a) Confusion matrix of MVGT-F on the SEED dataset. (b) Confusion matrix of MVGT-G on the SEED-IV dataset. Each row of the matrix represents the true labels while each column serves as the predicted labels.

Table II
ABLATION STUDY FOR THE CLASSIFICATION ACCURACIES (MEAN/STD) ON THE SEED AND SEED-IV DATASETS. SYMBOL "✓" INDICATES THE COMPONENT IS EMPLOYED.

Geometric Structure	Brain Region	SEED	SEED-IV
-	-	93.79/07.15	89.49/10.40
✓	-	94.52/06.04	90.00/09.62
-	✓	94.11/05.77	89.87/10.41
✓	✓	96.45/04.40	93.57/08.60

Fig. 3 illustrates the confusion matrices of MVGT-F on the SEED and MVGT-G on the SEED-IV, respectively. The values represent the classification accuracy of the model for different emotion classes. For the SEED dataset, our model achieves the highest accuracy in recognizing positive emotion (98.12%), followed by neutral emotion (96.38%), with negative emotion being slightly lower (94.73%). Only 0.33% of positive emotion samples are misclassified as negative, while only 0.76% of negative emotion samples are recognized as positive, indicating the model's effectiveness in distinguishing valence changes. For the SEED-IV dataset, our model performs best in recognizing neutral emotion, with an accuracy of 95.90%, while its performance on happy emotion is slightly lower than the other three emotions, with an accuracy of 90.76%. This could be attributed to the GENERAL scheme setting, making the model more sensitive to the balanced emotion.

Our model achieves state-of-the-art performance on both SEED and SEED-IV, primarily due to our comprehensive consideration of frequency, temporal, and spatial geometric information, combined with prior knowledge from neuroscience.

D. Ablation Study

To validate the effectiveness of spatial encodings, we conduct ablation experiments on the SEED and SEED-IV datasets, as presented in Table II. By removing both types of spatial encodings, we repeat the aforementioned experiments under the same experimental settings. On the SEED dataset, the model achieves an accuracy of 93.79% with a standard deviation of 7.15%. Compared to MVGT-F, the accuracy decreases by 2.66% and the standard deviation increases by 2.75% after removing spatial encodings. For the SEED-IV dataset, the accuracy drops by 4.08%, resulting in 89.49%, with the standard deviation rising by 1.80% to 10.40%, when compared to MVGT-G. The experiments demonstrate that incorporating spatial structure information benefits the model performance in emotion recognition tasks. Under experimental settings that consider only geometric structure or brain region structure, the model's classification accuracy improves over the plain model without any spatial encoding. Evidently, when considering both types of spatial structures simultaneously, the model performance significantly outperforms that of the plain model and models only using single spatial information. This indicates the effectiveness of our proposed spatial encodings and confirms that the expressive power of the graph transformer relies on the structural and positional encodings.

E. Visualization of Inter-channel relations

To better illustrate the correlations between channels, we visualize the inter-channel relations of MVGT-F on the SEED and MVGT-G on the SEED-IV. Given that the inter-channel relations might vary among different subjects, we calculate the average weights across all subjects. We focus on the last iteration of iterative refinement and select the 10 strongest connections of channel pairs. Fig. 4 shows the visualization results, where the rows represent the attention heads and the columns represent the layers of the MHA.

The parameters based on the SEED dataset indicate that emotion patterns are reflected in the activities of multiple brain regions. In the first layer of MVGT-F, the channels in the left frontal region have higher participation in the first attention head, while the channels in the right frontal region are more involved in the second head, potentially corresponding to positive and negative emotion patterns [9], respectively. In the second layer of the model, the parietal and occipital regions show higher involvement, which aligns with the findings on emotion patterns in [49]. As the model depth increases, the symmetrical connections in the lateral temporal regions of both hemispheres are enhanced, consistent with previous research by [5], [13], [16]. For the SEED-IV dataset, the connections in the frontal, parietal, and occipital regions are the most active, consistent with the findings of [13]. In the first attention

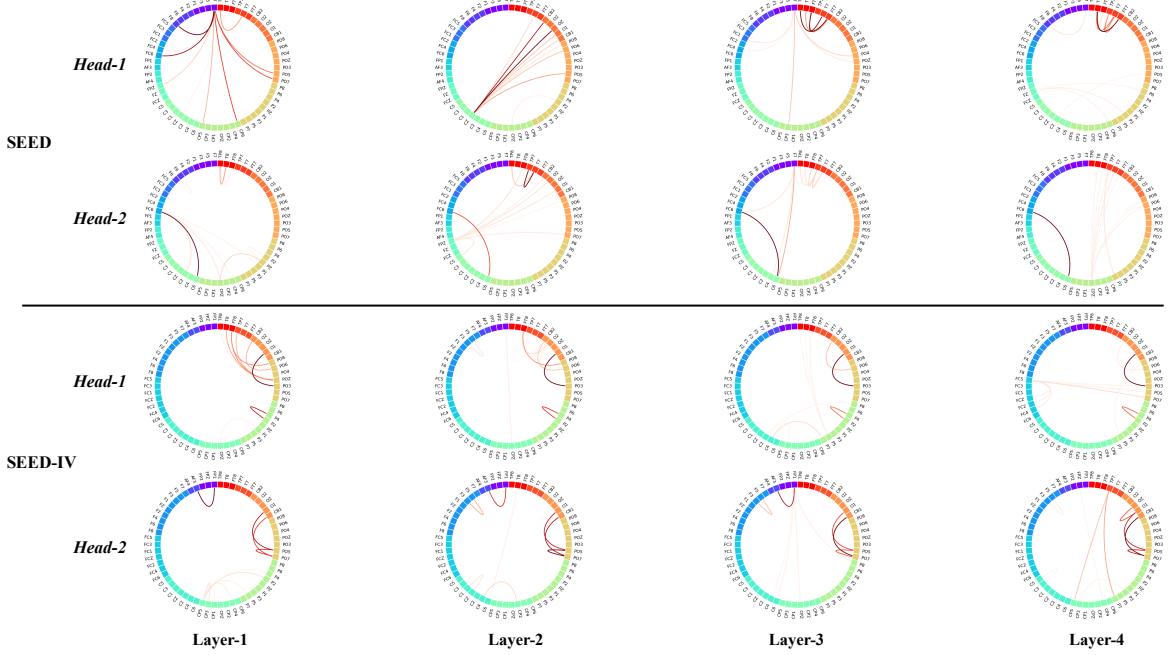


Figure 4. The learned inter-channel relationships from the SEED by the MVGT-F and from the SEED-IV by the MVGT-G are illustrated. The figures show the results of the last iteration in the iterative refinement, highlighting the top 10 channel pairs with the highest weights after softmax (darker colors indicate higher weights). Channels of the same brain region are represented in the same color. Rows correspond to attention heads, while columns represent the layers of the MHA mechanism.

head of MVGT-G, the strongest correlation is between O1 and PO3, followed by P4 and P2. Other connections are mainly distributed in the temporal and frontal regions. In the second head, the channel pairs (O1, PO5), (CB1, PO7), and (PO5, PO7) contribute the most to emotion recognition. Additionally, the connection between AF3 and FP1 provides important information for emotion processing, which aligns with the conclusions of [13], [16].

Overall, our model does not focus solely on the local information of a single brain region but instead considers both intra-regional and inter-regional information in depth. This confirms that emotional states result from interactions among widely distributed functional networks in the brain, as discussed by [50].

VI. CONCLUSIONS

In this paper, we propose a multi-view graph transformer based on spatial relations for EEG-based emotion recognition. Our model integrates information from multiple perspectives, including temporal, frequency and spatial domains. We incorporate spatial geometric encoding and brain region encoding to enhance the graph transformer’s ability to perceive spatial structures. Additionally, the model adaptively learns inter-channel relationships through the attention mechanism and the encoding of channel geometry. Extensive experiments on public emotion recognition datasets demonstrate that our proposed model outperforms other competitive baseline models.

Furthermore, the analysis of channel correlations indicates that emotional activities in the brain are not confined to a single local region but result from the coordinated action of multiple

brain areas. Information from frontal, parietal, occipital, and lateral temporal lobes is valuable for emotion recognition to varying extents.

REFERENCES

- [1] Babak A Taheri, Robert T Knight, and Rosemary L Smith. A dry electrode for eeg recording. *Electroencephalography and clinical neurophysiology*, 90(5):376–383, 1994.
- [2] Christopher Niemic. Studies of emotion: a theoretical and empirical review of psychophysiological studies of emotion. *Journal of Undergraduate Research*, 2004.
- [3] Lester I Goldfischer. Autocorrelation function and power spectral density of laser-produced speckle patterns. *Josa*, 55(3):247–253, 1965.
- [4] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pages 81–84. IEEE, 2013.
- [5] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.
- [6] Yisi Liu and Olga Sourina. Real-time fractal-based valence level recognition from eeg. In *Transactions on computational science XVIII: special issue on Cyberworlds*, pages 101–120. Springer, 2013.
- [7] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. Eeg-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, 2010.
- [8] Iris B Mauss and Michael D Robinson. Measures of emotion: A review. *Cognition and emotion*, pages 109–137, 2010.
- [9] Louis A Schmidt and Laurel J Trainor. Frontal brain electrical activity (eeg) distinguishes valence and intensity of musical emotions. *Cognition & Emotion*, 15(4):487–500, 2001.
- [10] Yang Li, Lei Wang, Wenming Zheng, Yuan Zong, Lei Qi, Zhen Cui, Tong Zhang, and Tengfei Song. A novel bi-hemispheric discrepancy model for eeg emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 13(2):354–367, 2020.

- [11] Yang Li, Wenming Zheng, Lei Wang, Yuan Zong, and Zhen Cui. From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*, 13(2):568–578, 2022.
- [12] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.
- [13] Peixiang Zhong, Di Wang, and Chunyan Miao. Egg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2020.
- [14] Rui Li, Yiting Wang, and Bao-Liang Lu. A multi-domain adaptive graph convolutional network for egg-based emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5565–5573, 2021.
- [15] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhan Zeng, and Cuntai Guan. Lggnnet: Learning from local-global-graph representations for brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2023.
- [16] Wei-Bang Jiang, Xu Yan, Wei-Long Zheng, and Bao-Liang Lu. Elastic graph transformer networks for eeg-based emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [17] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International conference on machine learning*, pages 1725–1735. PMLR, 2020.
- [18] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3538–3545, 2018.
- [19] Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.
- [20] Pablo Barceló, Egor V. Kostylev, Mikael Monet, Jorge Pérez, Juan Reutter, and Juan Pablo Silva. The logical expressiveness of graph neural networks. In *International Conference on Learning Representations*, 2020.
- [21] Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.
- [22] Michal Tepian et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.
- [23] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.
- [24] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.
- [25] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.
- [26] Luis Müller, Mikhail Galkin, Christopher Morris, and Ladislav Rampášek. Attending to graph transformers. *Transactions on Machine Learning Research*, 2024.
- [27] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.
- [28] Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34:21618–21629, 2021.
- [29] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.
- [30] Yu Shi, Shuxin Zheng, Guolin Ke, Yifei Shen, Jiacheng You, Jiyang He, Shengjie Luo, Chang Liu, Di He, and Tie-Yan Liu. Benchmarking graphomer on large-scale molecular modeling datasets. *arXiv preprint arXiv:2203.04810*, 2022.
- [31] Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu. A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6–14, 2022.
- [32] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR, 2019.
- [33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.
- [34] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- [35] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [37] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.
- [38] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.
- [39] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.
- [40] Hedy Kober, Lisa Feldman Barrett, Josh Joseph, Eliza Bliss-Moreau, Kristen Lindquist, and Tor D Wager. Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage*, 42(2):998–1031, 2008.
- [41] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Toumaj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.
- [42] Roland H Grabner and Bert De Smedt. Oscillatory eeg correlates of arithmetic strategies: a training study. *Frontiers in psychology*, 3:35080, 2012.
- [43] Muhammed Shuaibi, Adeesh Kolluru, Abhishek Das, Aditya Grover, Anuroop Sriram, Zachary Ulissi, and C Lawrence Zitnick. Rotation invariant graph neural networks using spin convolutions. *arXiv preprint arXiv:2106.09575*, 2021.
- [44] Tianle Cai, Shengjie Luo, Keyulu Xu, Di He, Tie-yan Liu, and Liwei Wang. Graphnorm: A principled approach to accelerating graph neural network training. In *International Conference on Machine Learning*, pages 1204–1215. PMLR, 2021.
- [45] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture, 2020.
- [46] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
- [47] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49(3):1110–1122, 2018.
- [48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [49] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Emotional state classification from eeg data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.
- [50] Lisa Feldman Barrett and Ajay Bhaskar Satpute. Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology*, 23(3):361–372, 2013.

Hypergraph Multi-modal Large Language Model: Exploiting EEG and Eye-tracking Modalities to Evaluate Heterogeneous Responses for Video Understanding

Minghui Wu^{1,2,3*} Chenxu Zhao^{2,3*} Anyang Su^{2,3*} Donglin Di³ Tianyu Fu³

Da An³ Min He² Ya Gao^{1,2} Meng Ma¹ Kun Yan^{1†} Ping Wang^{1†}

¹Peking University ²Mininglamp Technology ³Shanghai Artificial Intelligence Laboratory

{wuminghui, zhaochenxu, suanyang, hemin}@mininglamp.com, gaoya@stu.pku.edu.cn, {mameng, kyan2018, pwang}@pku.edu.cn

Abstract

Understanding of video creativity and content often varies among individuals, with differences in focal points and cognitive levels across different ages, experiences, and genders. There is currently a lack of research in this area, and most existing benchmarks suffer from several drawbacks: 1) a limited number of modalities and answers with restrictive length; 2) the content and scenarios within the videos are excessively monotonous, transmitting allegories and emotions that are overly simplistic. To bridge the gap to real-world applications, we introduce a large-scale Subjective Response Indicators for Advertisement Videos dataset, namely SRI-ADV. Specifically, we collected real changes in Electroencephalographic (EEG) and eye-tracking regions from different demographics while they viewed identical video content. Utilizing this multi-modal dataset, we developed tasks and protocols to analyze and evaluate the extent of cognitive understanding of video content among different users. Along with the dataset, we designed a Hypergraph Multi-modal Large Language Model (HMLLM) to explore the associations among different demographics, video elements, EEG and eye-tracking indicators. HMLLM could bridge semantic gaps across rich modalities and integrate information beyond different modalities to perform logical reasoning. Extensive experimental evaluations on SRI-ADV and other additional video-based generative performance benchmarks demonstrate the effectiveness of our method. The codes and dataset will be released at <https://github.com/mininglamp-MLLM/HMLLM>.

1. Introduction

With the advancement of Large Language Models (LLMs) [69] and Multi-modal Large Language Models [11, 31, 43,

*Authors contributed equally to this work.

†Corresponding authors.

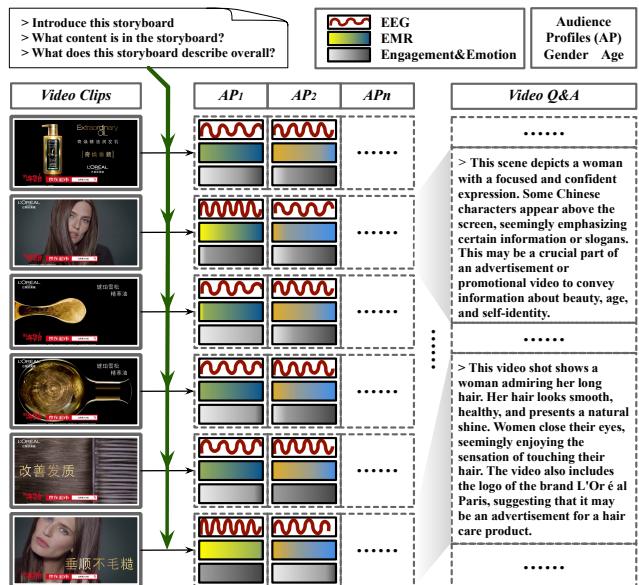


Figure 1. Our proposed Subjective Response Indicators for Advertisement Videos (SRI-ADV) dataset. Real-time signals captured by electroencephalographic (EEG) and eye-tracking devices reveal that Audience Profiles (AP) of varying genders and ages exhibit distinct engagements, emotions, and eye motion ratios (EMR) when exposed to various scenes and elements within the same advertisement video.

44], the field of video understanding has entered a new era. The advanced logical reasoning abilities of multi-modal LLMs facilitate a thorough analysis of explicit elements within videos. Moreover, these models can deduce the underlying implicit content of these explicit factors, leveraging the knowledge and experience acquired by LLMs. Existing benchmarks for video content question-and-answering, such as [30, 50, 72, 72, 76], provide a rich set of instruction labels. Alternatively, they exhibit several deficiencies as illustrated in Table 1: 1) the video content itself is overly simplistic, often only involving objective,

Table 1. Comparison of existing VideoQ&A datasets with ours (OE: open-ended, MC: multiple-choice, AP: Audience Profiles).

Datasets	Video source	Q&A generation	Q&A tasks	Modality	Videos	Q&A pairs	AvgAnsLen	MedScene
MSVD-QA [72]	MSVD	Auto	OE	Video	1,970	50,505	1.0	2
MSRVT-QA [72]	MSRVT	Auto	OE	Video	10,000	243,680	1.0	3
TGIF-QA [30]	TGIF	Auto&Human	OE & MC	Frame/Video	56,720	103,919	1.5	1
ActivityNet-QA [76]	ActivityNet	Human	OE	Video	5,800	58,000	1.3	7
Video-ChatGPT [50]	ActivityNet	Auto&Human	OE	Video	200	2,994	51.0	6
SRI-ADV-QA (ours)	Custom	Auto&Human	MC & OE	Video/EEG/EMR/AP	498	178,547	99.6	11

explicit factors, which does not support the exploration of deeper levels of video creativity and implicit factors. We utilize the **MedScene** metric to evaluate this issue, where **MedScene** denotes the median number of scene across all videos in the dataset. A higher number of scenes indicates greater complexity in video content; 2) the number of modalities included in these datasets are limited, generally confined to videos and frames; 3) the instruction labels concerning the length of answers are restricted to certain predetermined options, failing to assess the divergent and analytical abilities of LLMs. We utilize the **AvgAnsLen** to evaluate this issue, where **AvgAnsLen** represents the average text length of the answer portion across all Q&A pairs in the dataset. To address the issues mentioned above, we have prepared an extensive collection of content-rich advertisement videos, accompanied by a more comprehensive set of modality labels.

In the burgeoning field of cognitive neuroscience, the exploration of how individuals perceive and interpret video content has opened new avenues for understanding the intricate interplay between brain activity and media interaction [64]. Recent advancements in multi-modal data analysis have underscored the importance of leveraging diverse physiological signals to gain insights into the cognitive and emotional states of viewers [35]. Among these, Electroencephalographic (EEG) signals with their high temporal resolution, provide a direct measure of brain activity [56], capturing the nuanced and dynamic changes in cognitive states as individuals engage with video content. These signals embody the electrical manifestations of the brain’s complex neural dynamics, offering insights into the emotional and cognitive processes underpinning video content interpretation [55].

Inspired by the aforementioned context, we have utilized EEG and eye-tracking apparatus to collect and record the EEG and eye movement responses of individuals across various ages, genders, and professions while watching the same advertisement video. We aggregated this information into modality labels, introducing a novel, large-scale benchmark: the Subjective Response Indicators for Advertisement Videos dataset, namely **SRI-ADV**. As illustrated in Figure 1, our proposed dataset captures the subjective reactions of individuals watching videos through EEG and eye-tracking devices, fills the gaps in the video understand-

ing domain regarding the assessment of video appeal and implicit factors. How to effectively leveraging these multi-modal labels to uncover the latent associations among the modalities becomes the cornerstone for addressing deeper challenges in video understanding.

Graph-based methodologies exhibit superiority in exploring the associations among features, particularly hypergraphs, extending beyond traditional graph theory, offer a powerful framework for representing complex relationships in data [6]. In the context of video content analysis, hypergraphs can encapsulate the intricate associations among video elements, EEG signals, and eye-tracking data, allowing for the modeling of higher-order interactions that are not capturable through simple pairwise connections.

Utilizing the multi-modal information of the SRI-ADV dataset, coupled with the superiority of constructing associative features through hypergraph, we proposed a Hypergraph Multi-modal Large Language Model (**HMLLM**), integrating information from disparate modalities to perform logical reasoning and semantic analysis. By leveraging the rich information encoded in video content, along with EEG and eye-tracking data, HMLLM can bridge semantic gaps across modalities, offering a comprehensive understanding of the cognitive processes involved in video content interpretation.

The main contributions can be summarized as follows:

1. Introduction of a novel large-scale benchmark dataset: the Subjective Response Indicators for Advertisement Videos (SRI-ADV) dataset, a large-scale benchmark that captures real-time EEG and eye-tracking data from a diverse demographic while they watch advertisement videos. This dataset fills a significant gap in the field of video understanding by providing rich modality information and a comprehensive set of question-and-answer (Q&A) pairs that allow for the assessment of video creativity and implicit factors.

2. Development of the Hypergraph Multi-modal Large Language Model (HMLLM): we have developed a novel HMLLM that leverages the complex relationships among video elements, EEG signals, and eye-tracking data encapsulated in hypergraphs.

3. Extensive experimental evaluations demonstrating our method’s effectiveness: through rigorous experimental evaluations conducted on the SRI-ADV dataset and addi-

tional video Q&A datasets, we have demonstrated the effectiveness of our HMLLM.

2. Background

2.1. Video Understanding

Video understanding aims to create algorithms that allow machines to interpret videos with the same expertise as humans. Meanwhile, video emotion recognition [42][54][81] emphasizes the interplay between the emotions conveyed by the video and the viewer responses, collectively forming a critical component of video understanding. Most existing works focus on modeling objective and tangible visual properties of videos [16], particularly in action recognition [3, 7, 10, 17, 19, 20, 51, 58, 66, 68] and temporal action localization/detection [18, 46, 83]. However, the need for content recommendation systems has spurred research into subjective and intangible aspects (e.g. the appeal and memorability of content [14]), where various semantically rich information are considered [5, 13, 52, 82].

Compared with the above work, we present a new large-scale dataset filled with content-rich advertisement videos. This dataset includes a wider range of labels that cover both tangible and intangible aspects of content. Leveraging this dataset, we introduce an advanced hypergraph multi-modal large language model. This model is designed to simultaneously process various modalities, enabling it to conduct logical reasoning and perform in-depth semantic analysis of video content.

2.2. EEG-Based Emotion Recognition

Electroencephalography (EEG) signals provide detailed insights into brain activity related to emotions, offering spatial information on specific brain regions involved [8]. The Arousal-Valence model [57] is a key framework for classifying emotions along two dimensions. Xiaolin et al [60] explored various features to enhance the emotion recognition model. However, there's a shift towards deep learning due to the limitations of machine learning. The dynamical graph convolutional neural network (DGCNN) [59] was proposed to learn discriminative EEG features and interrelationships among EEG channels. Some works have moved towards multi-modal learning for robust results in EEG signal recognition tasks, such as integrating physiological signals in the multi-modal framework to enhance emotion recognition accuracy [71] and employing proper windowing and channel selection to avoid relying on the full length of EEG and EOG signals for classification [9]. Furthermore, advancements in neuromorphic computing led to the use of Spiking Neural Networks (SNN) [49] for classifying spatiotemporal EEG data with lower computational requirements [33].

2.3. Multi-modal Large Language Models

Multi-modal Large Language Models (MLLMs), primarily serving as vision-language models, transform images or videos into texts. These models are mainly divided into two categories: traditional large-scale pretraining [37, 38, 65] and instruction tuning using pre-trained LLMs [47, 75, 84]. The first category comprises models that blend a visual encoder with a language model, either developed from scratch or based on pre-existing models, possibly including a trainable module to bridge the two modalities. Utilizing autoregressive loss for text generation, these models are training on extensive image-text datasets, including image-text pairs [28, 37, 38, 65] and image-text sequence instances [2]. The second category, drawing inspiration from instruction-tuning techniques used in MLLMs [1, 53], incorporates instruction-following data to enhance MLLMs' zero- and few-shot learning abilities [15, 47, 75, 84]. A notable example is LLaVA [47], which employs a simple projection matrix to link a pre-trained visual encoder with an LLM, focusing initially on pre-training for feature alignment before comprehensive end-to-end fine-tuning. Some other works extend to video understanding by connecting video encoders to MLLMs [39, 45, 74, 78]. In addition to models that focus on combining images or videos with text, there are projects that incorporate even more types of data, like speech, audio, and sensor information [25, 61, 70, 77].

2.4. Hypergraph Learning

A hypergraph includes vertices and hyperedges, where hyperedges can connect multiple vertices. This structure is more adaptable and effective for representing complex relationships in data than traditional graphs [24]. Methods for creating hypergraphs fall into two groups: explicit and implicit. Explicit methods directly use the data structure to form hyperedges, like connecting vertices with shared attributes [27, 32]. Implicit methods, however, infer hyperedges from data without clear high-order links, utilizing approaches based on distance [22] or representations [48, 67]. Unlike static structures, some methods allow for hypergraph structure optimization, adjusting it during the learning phase. This involves adaptively changing weights on hyperedges [23] or sub-hypergraphs [80] to improve learning outcomes. Recent advancements have introduced deep hypergraph representation learning, a new approach that mainly divides into spectral [21, 73] and spatial [4, 26] categories based on how hypergraph convolution operator is defined.

3. SRI-RAV Dataset

In this section, we present the Subjective Response Indicators for Advertisement Videos (**SRI-ADV**) dataset. The SRI-ADV dataset not only focuses on the Objectivity Task

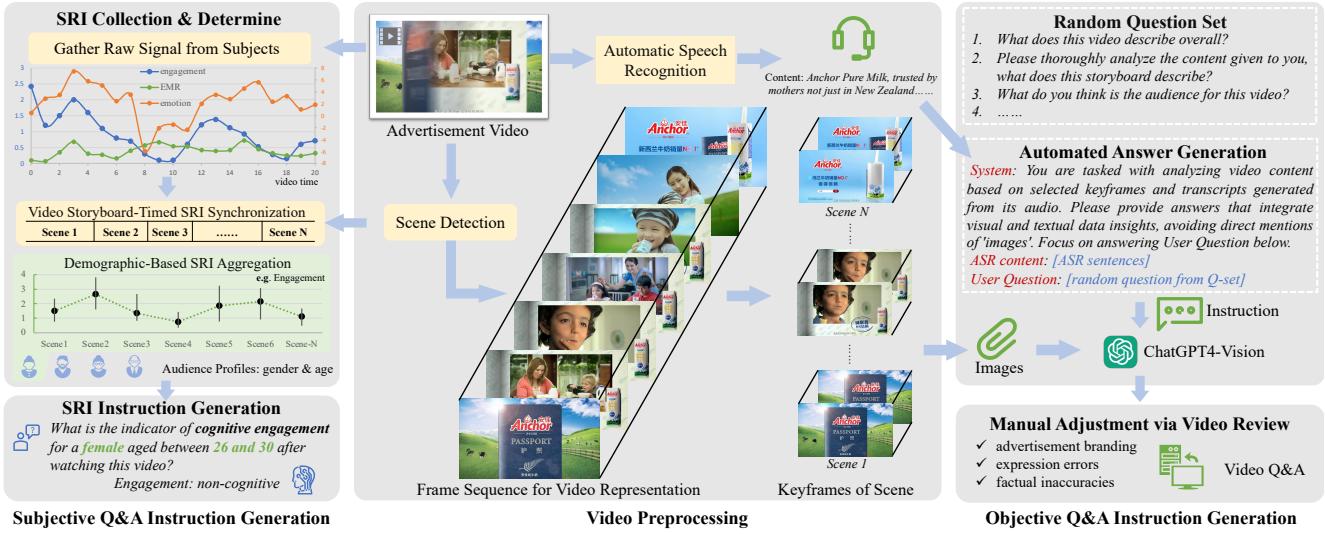


Figure 2. Generation pipeline of SRI-ADV dataset. The left side of this figure illustrates the process of SRI data collection, computation, and amalgamation. This involves acquiring raw signals from subjects, processing signals by video scenes, and pooling data from subjects with similar demographic profiles to obtain aggregated subjective response indicators and instruction for language models. The middle section depicts the video preprocessing with Frame Sequence for Video Representation (FSVR) by scene detection and Automatic Speech Recognition (ASR) for videos. On the right side, we present our proposed semi-automated video Q&A generation process, which leverages both video storyboarding from FSVR and dialogue text from ASR. This integration enriches video content comprehension, thereby facilitating both Subjectivity and Objectivity Tasks.

typically found in traditional video Q&A datasets but also meticulously collects subjective indices to enhance the richness. It encompasses a wide array of advertisement videos across different industries. To capture a diverse set of responses, we enlisted participants from various cities throughout Mainland China. These participants are equipped with EEG devices, enabling us to monitor their brainwave activities and eye motion ratios (EMR) in real-time while watching the advertisements. The collected data is subsequently analyzed to establish a benchmark for the classification of brainwave and EMR responses, which is elaborated in Sections 3.1 and 3.2.

Additionally, the SRI-ADV dataset includes an extensive video Q&A section to provide objective insights into the ads, facilitating model training and subjective index assessment. The task definition and protocol of our dataset are outlined in Section 3.3 and 3.4.

3.1. Frame Sequence for Video Representation

The SRI-ADV dataset features Chinese advertising videos from diverse fields such as food and beverages, household items, consumer electronics, cultural tourism, software, and automobiles. It comprises 498 curated landscape videos sourced from online platforms and TV commercial ads, each running for 15-30 seconds.

In this study, we introduce the Frame Sequence for Video Representation (**FSVR**) strategy to preprocess advertisement videos, as depicted in the middle part of Fig-

ure 2. We enhance the video scene sensitivity by integrating the AdaptiveDetector¹ for FSVR with specific parameters: adaptive_threshold = 2, min_scene_len = 10, window_width = 2. In the case of advertisement videos with frequent scene changes, the scene detection algorithm captures more information compared to average frame capture methods. Moreover, it is invaluable in minimizing redundant frames in videos primarily composed of static scenes.

By employing FSVR, we are able to deconstruct the temporal sequence of advertisement video frames, achieving capabilities including modality signal alignment, video content understanding, and semi-automated Q&A instruction generation.

3.2. Subjectivity: SRI Collection & Classification

We developed a sophisticated system for collecting subjective indicators. Each participant watches a series of advertisement videos using the device described in the appendix. During this process, we synchronously gather EEG and eye-tracking data, along with anonymized demographic details. Our study includes over 4,600 participants, ensuring a wide demographic representation. The diverse participant base spans white-collar workers, civil servants, students, and freelancers across various age groups and income brackets.

¹<https://www.scenedetect.com/>

The raw EEG signals are characterized by parameters such as $\alpha_1, \alpha_2 \dots \beta_2, \beta_3$ [34, 36], which is detailed in the appendix. Given the unique demands of advertisement video analysis, we pinpointed two pivotal EEG metrics: engagement and emotion, as delineated by Equation 1 and Equation 2, respectively.

$$EN_t = (\beta_2 + \beta_3) / (\alpha_3 + \alpha_2 + \beta_2 + \beta_3), \quad (1)$$

$$EM_t = (\alpha_3 - \alpha_2) / (\alpha_3 + \alpha_2) \times 100, \quad (2)$$

where EN_t and EM_t represent the engagement and emotion of the individual user at the sampling moment, respectively. Furthermore, we tracked eye movement data, defining the Eye Movement Ratio (EMR_t) as the proportion of time the participant's gaze fixates on the display relative to the total video duration.

The SRI Collection & Determine workflow, depicted on the left of Figure 2, captures sub-second high-frequency raw signals data. To align with video content's scene-based evolution, Video Storyboard-Timed SRI Synchronization was adopted, producing time-averaged and participant-specific SRIs. Demographic characteristics then grouped these SRIs into units of 5-20 same-gender participants with a maximum age difference of 5 years, such as {female, 120}, {male, 26-30}, and {female, 46-50}, as Demographic-Based SRI Aggregation in Equation 3.

$$\bar{X} = \frac{1}{P \cdot N} \sum_{i=1}^P \sum_{j=1}^N X_{p_i, t_j}, \forall p_i \in [AP], \forall t_j \in [t_1, t_2], \quad (3)$$

where X_{p_i, t_j} denotes the original Subjective Response Indicators such as EN_t , EM_t , and EMR_t . Each indicator associated with discrete values for participant p_i at specific timestamps t_j , where t_j signifies the effective sampling moment instances within the video storyboard timeframe from FSVR in Section 3.1.

For quantitative analysis, we meticulously examined data distribution across various Audience Profile segments. Engagement was categorized into two groups using the Leuven Engagement Scale (LES) and its distribution. Emotion and EMR indicators, which followed normal distributions, were divided into three equal categories. For detailed data distribution, refer to the appendix. The SRI Instruction Generation protocol is detailed in Table 2.

3.3. Objectivity: Semi-automated Generation

In addition to subjective indicators from Audience Profiles, we developed a semi-automated annotation pipeline for ChatGPT4-Vision (GPT4V) to obtain Objective Video Q&A, depicted in Figure 2. Although GPT4V cannot process videos, it supports multiple consecutive key-frames simultaneously. Based on FSVR in Video Preprocessing, we extracted middle frames from each shot as key-frames that

Table 2. Task and Protocol of SRI-ADV Dataset. In Task1, Protocol1 (**P1**) targets a broad audience. Protocol2 (**P2**), based on P1, contains SRI to Audience Profiles.

Task Name	1. Subjectivity	2. Objectivity
Eva. Form	Multi-classification	Text generation
Train Video	426	426
Test Video	72	72
Train Q&A	145,107	5762
Test Protocol	P1	P2
Test Q&A	2,640	26,724
		954

effectively represent the entire video. During each invocation of GPT4V to automatically generate answers, questions are selected randomly from the Random Question Set to enhance the diversity of Q&A sessions, along with providing ASR text and FSVR key-frames. Lastly, annotators were carefully selected to manually refine objective Q&A instruction from Automated Answer Generation, addressing issues like advertisement branding, expression errors, and factual inaccuracies.

3.4. Data Overview, Tasks and Protocols

Based on the processing presented in Sections 3.2 and 3.3, SRI-ADV is categorized into subjectivity and objectivity tasks. The subjectivity task examines the SRI, whereas the objectivity task is dedicated to the qualitative analysis of video content and audience perception. As shown in Table 2, we present the tasks, protocols, and instructions associated with the SRI-ADV dataset.

Task 1, entitled **Subjectivity**, is formulated as a classification task, aimed at examining the influence of video content and user characteristics on the SRI. We developed two experimental protocols to guide this investigation. The first protocol (**P1**) is designed to assess the SRI ability of a broad audience, involving the analysis of average responses across different videos. This approach is relatively straightforward. The second protocol (**P2**) introduces a layer of complexity by focusing on the SRI discernment of particular user demographics. This necessitates a comprehensive examination of how response patterns fluctuate among diverse user cohorts.

Task 2, designated as **Objectivity**, mirrors the video Q&A tasks prevalent in prior datasets, as described in Section 3.3. Building on the method outlined in [50], this study conducts a supervised analysis of the answers generated, assessing their accuracy and allocating scores. This approach is designed to objectively ascertain the narrative coherence of the advertisement content and its efficacy in captivating the target audiences.

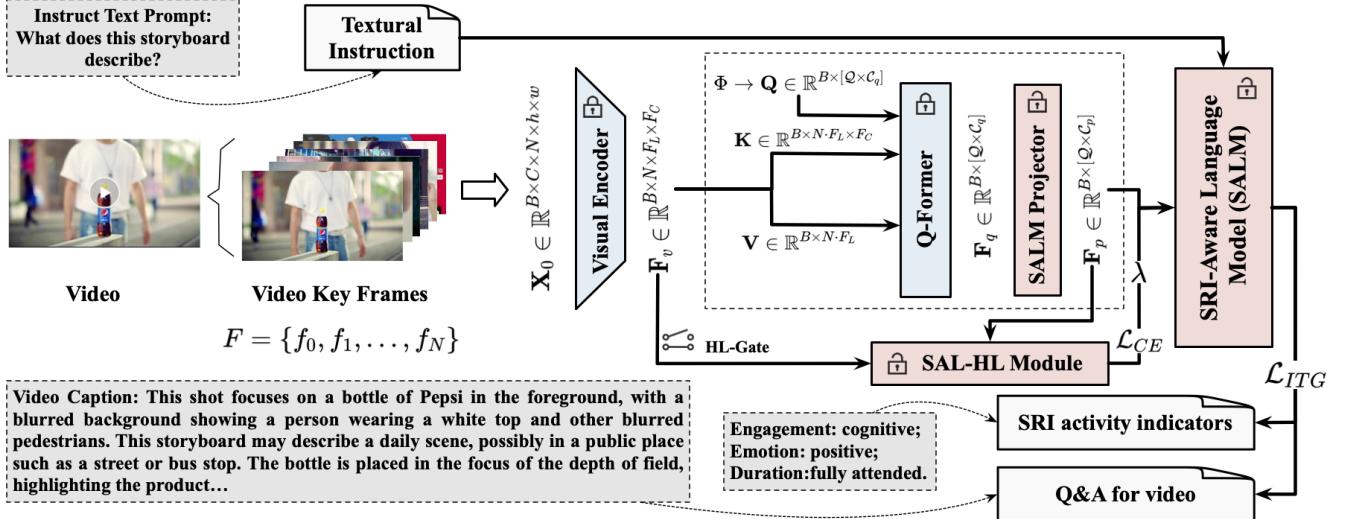


Figure 3. Overview of the Hypergraph Multi-modal Large Language Model (HMLLM). The architecture comprises a suite of pre-trained models, including a “Visual Encoder”, “Q-Former”, and the “SRI-Aware Language Model (SALM)”, which are initially frozen and subsequently fine-tuned through strategic training procedures. More importantly, our model incorporates a designed “SRI-Aware Language Hypergraph Learning (SAL-HL)” module that is trained de novo via a combined loss function. During inference, the HMLLM generates SRI and Q&A responses tailored to the video content, thereby providing a deeper level of engagement and comprehension.

4. Method

This section elaborates on the Hypergraph Multi-modal Large Language Model (HMLLM), an approach designed to intelligently process video clips and textual prompts for generating contextually relevant text, including Subjective Response Indicators (SRI). Central to our methodology are several key components as depicted in Figure 3: Visual Encoder, Query Former (Q-Former), SALM Projector, SRI-Aware Language Model (SALM), and SAL-HL Module. All components mentioned above synergistically orchestrated across two primary phases: SALM Warm-Up and SAL-HL Fine-Tuning, as depicted in our model architecture (refer to Figure 3). The pseudocode in the appendix illustrates the detailed training process.

4.1. SALM Warm Up

We begin by detailing the initial stage. The approach ingests brief video clips and corresponding textual prompts, extracting key frames from the videos using a predefined, static extraction strategy, which can be either random or uniformly distributed. These key frames are represented as $F = \{f_0, f_1, \dots, f_N\}$, with N signifying the number of extracted frames. These key frames are then pre-processed to form the initial data matrix, denoted by $\mathbf{X}_0 \in \mathbb{R}^{B \times C \times N \times h \times w}$, where B, C, N, h , and w correspond to the batch size, color channels (RGB), the number of keyframes, and the resized dimensions of the frames, respectively. The initial data matrix \mathbf{X}_0 is fed into a pre-trained visual encoder to yield initial visual representations, expressed as

$\mathbf{F}_v \in \mathbb{R}^{B \times N \times F_L \times F_C}$, with F_L and F_C representing the length and channels of features, respectively.

During the first training phase, the “Hypergraph Learning Gate (HL-Gate)” remains inactive while the Q-Former and SALM are warmed up. The visual features \mathbf{F}_v are then input into the frozen Q-Former as the Key ($\mathbf{K} \in \mathbb{R}^{B \times (N \cdot F_L) \times F_C}$) and Value ($\mathbf{V} \in \mathbb{R}^{B \times (N \cdot F_L)}$) for the attention mechanism. The Query in the Q-Former is initialized as either a random or null set, represented by $\mathbf{Q} \in \mathbb{R}^{B \times (\mathcal{Q} \times \mathcal{C}_q)}$, where $\mathcal{Q} \times \mathcal{C}_q$ are the predefined hyperparameters for the length and channels of the query. Subsequently, we introduce an “SALM Projector”, a multi-layer perceptron that follows the Q-Former, capable of reshaping the data and introducing additional learning parameters into the model. The output of projector is denoted as $\mathbf{F}_p \in \mathbb{R}^{B \times (\mathcal{Q} \times \mathcal{C}_p)}$, with \mathcal{C}_p being another predefined hyperparameter. The SRI-Aware Language Model (SALM) is then engaged, taking the output of the SALM Projector (\mathbf{F}_p) and the corresponding textual instructions as inputs during the initial warm-up training stage. The SALM is trained using the Image-grounded Text Generation (ITG) loss function [38] (\mathcal{L}_{ITG}), which instructs the Q-Former to generate text conditioned on the input images. The goal of the ITG loss is to minimize the difference between the generated caption $\tilde{\mathcal{Y}}_{qa} \leftarrow \text{SALM}(\mathbf{F}_p, T)$ and the ground-truth caption \mathbf{Y}_{gt} . This is typically achieved using a cross-entropy loss computed over the words or tokens in the caption. The

ITG loss function can be mathematically represented as:

$$\mathcal{L}_{ITG} = - \sum \left(\log \mathbb{P}(\mathbf{Y}_{gt_i} | \mathbf{Y}_{gt_1}, \dots, \mathbf{Y}_{gt_{i-1}}, \mathbf{F}_v) \right) \quad (4)$$

where $\mathbb{P}(\mathbf{Y}_{gt_i} | \mathbf{Y}_{gt_1}, \dots, \mathbf{Y}_{gt_{i-1}}, \mathbf{F}_v)$ denotes the probability of generating the i -th word in the caption given the previous words and the visual features extracted from the image. The summation encompasses all words or tokens in the ground-truth caption.

In our approach, we integrate specific strategies from BLIP2 [38] to address the limitation of Q-Former architecture on direct interactions between the image encoder and text tokens. Following the aforementioned training procedure, the SALM Projector and SALM are adequately warmed up, preparing them for subsequent fine-tuning optimization.

4.2. SAL-HL Fine-tune

In the subsequent fine-tuning phase, the hypergraph learning gate (HL-Gate) is activated, and the hypergraph learning module (SAL-HL) undergoes training in tandem with the fine-tuning of the SRI-Aware Language Model (SALM). As delineated in Figure 3, the SAL-HL module receives the initial visual features (\mathbf{F}_v) and the representations of the projected frames (\mathbf{F}_p) produced by the warmed SALM Projector as inputs.

The SAL-HL module initiates the process by merging these two feature sets (i.e., $\mathbf{F}_p, \mathbf{F}_v$) and then pooling them to generate frame-level representations (\mathbf{F}_{frame_level}). This process is formulated as:

$$\mathbf{F}_{frame_level} = \text{Pool}(\text{Feature_Mixer}(\mathbf{F}_p, \mathbf{F}_v)). \quad (5)$$

The *Feature_Mixer* denotes the mixing operation between two feature matrices, which can be implemented as a multi-layer perceptron (MLP). Each frame, denoted as f_i for $i \in [0, N]$, is considered a vertex (\mathcal{V}) within the hypergraph structure (\mathcal{G}), which facilitates the establishment of high-order relationships among the frames. The construction of the hypergraph entails the application of a clustering algorithm that links frames with similar latent visual features. After constructing the hypergraph, we proceed to train the Hypergraph Neural Network (HGNN) [21] in parallel with the Structured Attention Layer Mechanism (SALM). This process is mathematically formulated as follows:

$$\tilde{\mathcal{Y}}_{sri} = \sigma \left(\mathbf{D}_v^{-1/2} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^\top \mathbf{D}_v^{-1/2} \cdot \mathbf{F}_{frame_level} \cdot \Theta \right), \quad (6)$$

where $\tilde{\mathcal{Y}}_{sri}$ represents the predicted output from the SALM-enhanced HGNN, and σ denotes a non-linear activation function, which introduces the necessary non-linearity into the model for capturing complex patterns. $\mathbf{D}_e \in \mathbb{R}^{E \times E}$,

$\mathbf{D}_v \in \mathbb{R}^{N \times N}$, and $\mathbf{W} \in \mathbb{R}^{E \times E}$ denote the diagonal degree matrix of hyperedges, the degree matrix of vertices, and weight matrix of hyperedges, respectively. $\mathbf{H} \in \mathbb{R}^{N \times E}$ signifies the incidence matrix that connects hyperedges to their constituent vertices. $\sigma(\cdot)$ denotes the nonlinear activation function (e.g., LeakyReLU(\cdot)). Θ is a diagonal matrix representing the learnable parameters updated by the *Cross_Entropy* loss function in the fine-tuning loop. It functions similarly to a multilayer perceptron (MLP) layer. Finally, \mathbf{F}_{frame_level} represents the input feature vectors associated with the vertices of the hypergraph. By employing this formulation, we effectively leverage the structural complexity of the hypergraph to enhance the learning capabilities of the HGNN, enabling it to capture and utilize the intricate relationships inherent within the data. This joint training regimen integrates two loss functions: the Cross-Entropy loss (\mathcal{L}_{CE}) and the Image-grounded Text Generation (ITG) loss from the prior stage. The combined loss function is expressed as:

$$\mathcal{L} = \mathcal{L}_{ITG} + \lambda \cdot \mathcal{L}_{CE}, \quad (7)$$

where λ is a hyperparameter that balances the influence of the Cross-Entropy loss and the ITG loss on the overall optimization process. This composite loss function ensures that the model not only generates text that is grounded in the visual content but also adheres to the learned high-order relationships within the hypergraph structure. This enhances the model's capability to capture intricate interactions and dependencies among video frames.

5. Experiment

Metrics. In our study, the Subjectivity Task of SRI-ADV is structured in a multiple-choice question (**MC**) format. To evaluate its performance, we employ Accuracy (**Acc**) and F1 score as our metrics. For the zero-shot evaluation of the Subjectivity Task, we have devised a unique prompt, with detailed information provided in the appendix. The subsequent task, named Objectivity, involves open-ended (**OE**) text generation. For its evaluation, we introduce evaluation measures [50] based on assessments using GPT-3.5 Turbo.

Implementation Details. We employ UMT-L [41] as the visual encoder and Vicuna-7B-v0 [12] as the base model for the SRI-Aware Language Model (SALM). Following the methodology of BLIP2 [38], QFormer is configured with 64 queries. Throughout both the inference and training phases, we adhere to the **FSVR** strategy detailed in Section 3.1, which involves representing each video with 8 key frames. Further details of hypergraph construction can be found in the appendix.

Table 3. Results of different models on Subjectivity task (Engagement, Emotion, and EMR Duration). Using the Frame Sequence for Video Representation (FSVR) strategy is denoted by a “ Δ ”.

Models	Protocol	Settings	Engagement (2 classes)		Emotion (3 classes)		EMR Duration (3 classes)	
			Acc	F1	Acc	F1	Acc	F1
Random	P1	—	50.44	49.93	32.30	26.26	35.01	32.10
	P2	—	50.14	50.00	33.13	33.03	33.52	33.18
GPT4VΔ [1]	P1	Zero-shot	58.57	71.95	52.46	50.67	49.94	53.43
	P2	Zero-shot	45.62	61.53	36.40	43.65	39.39	47.04
Gemini-pro-visionΔ [63]	P1	Zero-shot	59.89	73.70	17.66	20.00	46.40	47.96
	P2	Zero-shot	46.16	63.31	30.56	43.10	36.20	43.96
Video-LLaVA [45]	P1	Zero-shot	60.06	74.50	61.39	71.30	45.26	57.48
	P2	Zero-shot	46.38	61.38	31.04	42.71	31.56	49.30
Video-LLaVA [45]	P1	Finetune	66.29	66.85	72.33	81.94	61.05	61.80
	P2	Finetune	52.58	52.69	38.62	44.72	41.28	50.84
Video-Chat2 [40]	P1	Finetune	75.34	76.95	71.36	75.78	57.39	60.80
	P2	Finetune	60.06	60.02	39.66	40.24	44.06	45.51
HMLLM (Ours)	P1	Finetune	78.41	79.26	78.41	84.83	62.05	62.43
	P2	Finetune	64.43	64.65	43.20	48.84	51.96	56.24

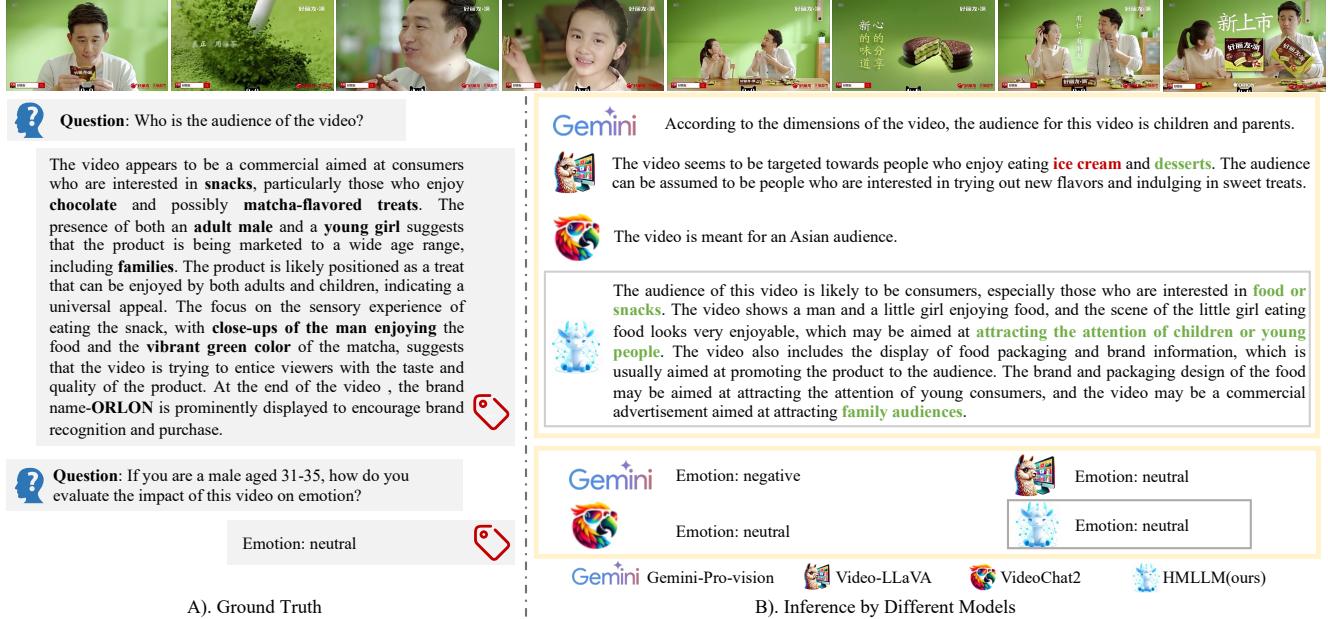


Figure 4. Qualitative analysis of SRI-ADV. Green signifies accurate descriptions, while red denotes incorrect responses.

5.1. Intra Testing

5.1.1 Intra Testing on SRI-ADV

As described in Section 3.4, we have designed two tasks on the collected SRI-ADV dataset, namely Subjectivity and Objectivity.

Subjectivity task. In this task, two protocols are encompassed, *i.e.*, P1, and P2. As shown in Table 3, we present our proposed method HMLLM alongside those from renowned MLLMs such as Gemini-pro-vision, GPT4V, Video-LLaVA, and Video-Chat2. The results, observed

from top to bottom, indicate:

1) For zero-shot inference, GPT4V and Video-LLaVA significantly outperformed the random across-the-board. Gemini-pro-vision underperformed the Random baseline in predicting the accuracy of Engagement and Emotion. Video-Chat2’s failure to follow instructions made it difficult to obtain meaningful results. The settings and prompts of zero-shot inference can be found in the appendix.

2) Upon fine-tuning the models with the SRI-ADV dataset, we observed notable improvements in performance for both Video-LLaVA and Video-Chat2 across both P1 and P2, compared to their initial zero-shot configurations.

Table 4. Comparative performance of different models on the Objectivity task. Using the FSVR strategy is denoted by a ” \triangle ”. The underline of GPT4V denotes the upper bound. We compute the Accuracy (Acc) and VideoChatGPT-Score (Score) [50] of the proposed method HMMLM and other compared state-of-the-art methods on testing data.

Models	Settings	Acc	Score [50]
GPT4V \triangle	Zero-shot	<u>84.80</u>	<u>3.99</u>
Gemini-pro-vision \triangle	Zero-shot	27.15	2.35
Video-LLaVA [45]	Zero-shot	15.20	2.06
Video-Chat2 [40]	Zero-shot	21.80	2.11
Video-LLaVA [45]	Finetune	44.76	3.03
Video-Chat2 [40]	Finetune	49.27	3.12
HMMLM (Ours)	Finetune	50.52	3.13

Table 5. Results of video conversation benchmark [50]. CI: Correctness of Information, DO: Detail Orientation, CU: Contextual Understanding, TU: Temporal Understanding, C: Consistency.

Models	CI	DO	CU	TU	C	Avg.
Video LLaMA [78]	1.96	2.18	2.16	1.82	1.79	1.98
Video Chat [39]	2.23	2.50	2.53	1.94	2.24	2.29
LLaMA Adapter [79]	2.03	2.32	2.30	1.98	2.15	2.16
Video-ChatGPT [50]	2.40	2.52	2.62	1.98	2.37	2.38
Video-Chat2 [40]	3.02	2.88	3.51	2.66	2.81	2.98
HMMLM (Ours)	3.12	2.86	3.52	2.61	2.91	2.99

Table 6. Results of λ on Protocol2 of the Subjectivity Task.

λ	Engagement		Emotion		EMR	
	ACC	F1	ACC	F1	ACC	F1
0.0	60.06	60.02	39.66	40.24	44.06	45.51
0.05	62.69	63.66	42.62	42.46	52.22	54.78
0.1	64.43	64.65	43.20	48.84	51.96	56.24
0.2	62.38	63.28	38.06	39.79	50.57	50.56
0.5	60.86	60.83	40.17	42.34	50.89	52.47

Moreover, our proposed HMMLM demonstrated the most substantial advancements, consistently outperforming other state-of-the-art methods across all evaluated metrics and protocols.

In detail, within Protocol 1, HMMLM surpassed the leading benchmarks in the categories of Engagement (2 classes), Emotion (3 classes), and EMR Duration (3 classes). The improvements were remarkable, showing enhancements in (accuracy, F1) scores by (3.07, 2.31), (6.08, 2.89), and (1.00, 0.63), respectively. These results underscored the efficacy of our method in accurately capturing and analyzing both engagement and emotional dynamics, as well as predicting EMR duration with high precision. For Protocol 2, the superiority of HMMLM is equally evident. Again, it outshone the best-existing benchmarks in Engagement (2 classes), Emotion (3 classes), and EMR Duration (3 classes), with enhancements in (accuracy, F1) scores by

(4.37, 1.34), (3.54, 4.12), and (7.90, 5.40), respectively. These findings highlight the robustness and adaptability of our model across different protocols, further establishing its potential for widespread applicability in real-world scenarios.

Objectivity Task. In the exploration of the objectivity task, as detailed in Section 3.3, we meticulously refined the ground truth (GT) by manually correcting annotations initially provided by GPT4V. This meticulous process contributed to the notably high zero-shot inference capabilities observed for GPT4V. Given that Gemini-pro-vision and GPT4V inherently lack support for video inputs, we integrated Frame Sequence Video Representation (FSVR) technology to bridge this gap. This adaptation endowed both models with the ability to process video inputs, thus expanding their applicability across a wider range of tasks. As shown in Table 4, GTP4V became the upper bound in a zero-shot setting because we semi-automatically utilized it for labeling, as described in Section 3.3. When the narrative shifts upon the fine-tuning of our models with the SRI-ADV dataset. Both Video-LLaVA and Video-Chat2 showcased substantial enhancements in their performance metrics, surpassing their initial zero-shot configurations. This improvement highlights the transformative impact of targeted training on model efficacy. Notably, our proposed HMMLM method emerged as a formidable contender, eclipsing other models in performance across the board. Specifically, HMMLM outperformed the best baseline, Video-Chat2, in terms of accuracy and the Score [50] by 1.25 and 0.01, respectively.

The results not only validate the effectiveness of fine-tuning with the SRI-ADV dataset but also emphasize that our HMMLM method sets a new benchmark in model performance.

5.1.2 Intra Testing on Video Conversation Benchmark

To further validate the performance of HMMLM, we conducted experiments on other video-based generative performance benchmarks. Following the setup of Video-ChatGPT[50], we present the performance of our proposed HMMLM, detailed in the last row of Table 5. Experimental results demonstrate that the HMMLM effectively enhances both Contextual Understanding and Consistency. Given the HMMLM did not overemphasize temporal details, a slight decrease in Temporal Understanding was observed.

5.2 Ablation Study

Effect of λ . In the course of training HMMLM, a series of ablation studies were carried out on the λ in Equation 7, the results of which are detailed in Table 6. The integration of the SAL-HL Module significantly bolstered the model’s proficiency in capturing subjective metrics, culmi-

nating in optimal performance at a λ value of 0.1. Beyond this threshold, any further increase in λ resulted in a slight decrease in performance, likely due to an overemphasis on SAL-HL features at the expense of the SALM’s inferential capabilities. Despite this, HMMLM consistently surpasses the baseline model ($\lambda = 0.0$) in terms of inference strength, demonstrating the beneficial impact of the hypergraph integration on the model’s overall performance.

5.3. Analysis and Visualization

We further present a qualitative comparison in Figure 6. HMMLM demonstrates an enhanced ability to generate longer and more comprehensive responses for Objectivity Tasks. This improvement can be attributed to the longer average context length of our dataset, which facilitates a deeper understanding of video content by enabling detailed analysis of advertising plots and visual elements. More detailed qualitative analyses are available in the appendix.

6. Conclusion

In this paper, we released a large-scale SRI-ADV dataset with two challenging tasks. We hope it will push cutting-edge research in video understanding. Besides, we proposed a novel HMMLM approach that enhances the language model by constructing a hypergraph feature space across modalities, thereby providing semantically richer associative features. Finally, we conducted a comprehensive set of experiments on both SRI-ADV and other video-based generative datasets, verifying the significance of the proposed dataset and method.

7. Acknowledgments

This work was supported by the Brain-like General Vision Model and Applications project (Grant No. 2022ZD0160403), China Postdoctoral Science Foundation (2023M740079, GZC20230058).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 8
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3
- [4] Devanshu Arya, Deepak K Gupta, Stevan Rudinac, and Marcel Worring. Hypersage: Generalizing inductive representation learning on hypergraphs. *arXiv preprint arXiv:2010.04558*, 2020. 3
- [5] David Azcona, Enric Moreu, Feiyan Hu, Tomás E Ward, and Alan F Smeaton. Predicting media memorability using ensemble models. *CEUR Workshop Proceedings*, 2020. 3
- [6] Claude Berge. *Hypergraphs: combinatorics of finite sets*. Elsevier, 1984. 2
- [7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 3
- [8] Gyorgy Buzsaki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304(5679):1926–1929, 2004. 3
- [9] Huili Cai, Xiaofeng Liu, Rongrong Ni, Siyang Song, and Angelo Cangelosi. Emotion recognition through combining eeg and eog over relevant channels with optimal windowing. *IEEE Transactions on Human-Machine Systems*, 2023. 3
- [10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3
- [11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1
- [12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 7
- [13] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2531–2540, 2019. 3

- [14] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys (CSUR)*, 52(2):1–37, 2019. 3
- [15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructclip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [16] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 593–610. Springer, 2020. 3
- [17] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 3
- [18] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 768–784. Springer, 2016. 3
- [19] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 3
- [20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3
- [21] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3558–3565, 2019. 3, 7
- [22] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. 3-d object retrieval and recognition with hypergraph analysis. *IEEE transactions on image processing*, 21(9):4290–4303, 2012. 3
- [23] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing*, 22(1):363–376, 2012. 3
- [24] Yue Gao, Zizhao Zhang, Haojie Lin, Xibin Zhao, Shaoyi Du, and Changqing Zou. Hypergraph learning: Methods and practices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2548–2566, 2020. 3
- [25] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3
- [26] Jing Huang and Jie Yang. Unignn: a unified framework for graph and hypergraph neural networks. *arXiv preprint arXiv:2105.00956*, 2021. 3
- [27] Sheng Huang, Mohamed Elhoseiny, Ahmed Elgammal, and Dan Yang. Learning hypergraph-regularized attribute predictors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–417, 2015. 3
- [28] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [29] HuEdward J., Yulong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv: Computation and Language,arXiv: Computation and Language*, 2021. 15
- [30] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2
- [31] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1
- [32] Cliff Joslyn, Sinan Aksoy, Dustin Arendt, Louis Jenkins, Brenda Praggastis, Emilie Purvine, and Marcin Zalewski. High performance hypergraph analytics of domain name system relationships. In *HICSS 2019 symposium on cybersecurity big data analytics*, 2019. 3
- [33] Nikola Kasabov and Elisa Capecci. Spiking neural network methodology for modelling, classification and understanding of eeg spatio-temporal data measuring cognitive processes. *Information Sciences*, 294:565–575, 2015. 3
- [34] Chamandeep Kaur, Preeti Singh, et al. Eeg derived neuronal dynamics during meditation: Progress and challenges. *Advances in preventive medicine*, 2015, 2015. 5, 14
- [35] Ashmit Khandelwal, Aditya Agrawal, Aanisha Bhattacharyya, Yaman K Singla, Somesh Singh, Uttaran Bhattacharya, Ishita Dasgupta, Stefano Petrangeli, Rajiv Ratn Shah, Changyou Chen, et al. Large content and behavior models to understand, simulate, and optimize content and behavior. *arXiv preprint arXiv:2309.00359*, 2023. 2
- [36] Wolfgang Klimesch. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(2-3):169–195, 1999. 5, 14
- [37] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2, 2023. 3
- [38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

- frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3, 6, 7
- [39] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3, 9
- [40] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multimodal video understanding benchmark, 2023. 8, 9, 15
- [41] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models, 2023. 7
- [42] Mingcheng Li, Dingkang Yang, and Lihua Zhang. Towards robust multimodal sentiment analysis under uncertain signal missing. *IEEE Signal Processing Letters*, 2023. 3
- [43] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXXI 16*, pages 121–137. Springer, 2020. 1
- [44] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1
- [45] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 3, 8, 9
- [46] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 988–996, 2017. 3
- [47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3
- [48] Qingshan Liu, Yubao Sun, Cantian Wang, Tongliang Liu, and Dacheng Tao. Elastic net hypergraph learning for image clustering and semi-supervised classification. *IEEE Transactions on Image Processing*, 26(1):452–463, 2016. 3
- [49] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997. 3
- [50] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*, 2023. 1, 2, 5, 7, 9
- [51] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021. 3
- [52] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 223–240. Springer, 2020. 3
- [53] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3
- [54] Jicai Pan, Shangfei Wang, and et. al. Representation learning through multimodal attention and time-sync comments for affective video content analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 42–50, 2022. 3
- [55] Edmund T Rolls. *The brain, emotion, and depression*. Oxford University Press, 2018. 2
- [56] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019. 2
- [57] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980. 3
- [58] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 3
- [59] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2020. 3
- [60] Xiaolin Song, Qiaoju Kang, Zekun Tian, Yi Yang, Sihao Yang, Qiang Gao, and Yu Song. Eeg-based emotion classification with wavelet entropy feature. In *2020 Chinese Automation Congress (CAC)*, pages 5685–5689, 2020. 3
- [61] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 3
- [62] Abdulhamit Subasi, Turker Tuncer, Sengul Dogan, Dahiru Tanko, and Unal Sakoglu. Eeg-based emotion recognition using tunable q wavelet transform and rotation forest ensemble classifier. *Biomedical Signal Processing and Control*, 68:102648, 2021. 14
- [63] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 8
- [64] S Vajda. The mathematical theory of communication. by claude e. shannon and warren weaver. pp. 117. 1949. *The Mathematical Gazette*, 34(310):312–313, 1950. 2
- [65] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 3
- [66] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition.

- In *European conference on computer vision*, pages 20–36. Springer, 2016. 3
- [67] Meng Wang, Xueliang Liu, and Xindong Wu. Visual classification by hypergraph modeling. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2564–2574, 2015. 3
- [68] Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang, and Shanshan Hao. I3d-lstm: A new model for human action recognition. In *IOP conference series: materials science and engineering*, page 032035. IOP Publishing, 2019. 3
- [69] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1
- [70] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 3
- [71] Xun Wu, Wei-Long Zheng, Ziyi Li, and Bao-Liang Lu. Investigating eeg-based functional connectivity patterns for multimodal emotion recognition. *Journal of neural engineering*, 19(1):016012, 2022. 3
- [72] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yuetong Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 1, 2
- [73] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. *Advances in neural information processing systems*, 32, 2019. 3
- [74] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3
- [75] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3
- [76] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuetong Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 1, 2
- [77] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023. 3
- [78] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3, 9
- [79] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Ao-jun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 9
- [80] Zizhao Zhang, Haojie Lin, Xibin Zhao, Rongrong Ji, and Yue Gao. Inductive multi-hypergraph learning and its application on view-based 3d object classification. *IEEE Transactions on Image Processing*, 27(12):5957–5968, 2018. 3
- [81] Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *CVPR 2023*, pages 18888–18897, 2023. 3
- [82] Tony Zhao, Irving Fang, Jeffrey Kim, and Gerald Friedland. Multi-modal ensemble models for predicting video memorability. *arXiv preprint arXiv:2102.01173*, 2021. 3
- [83] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaonan Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2914–2923, 2017. 3
- [84] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

A. Analysis of EEG Raw Signal

Electroencephalography (EEG) stands as a pivotal method for recording the electrical activity of the brain. This is achieved through the placement of electrodes across the scalp, shown in Figure 5, to detect electrical signals from neurons. These signals are instrumental in delineating the brain's activity patterns across various cognitive states, providing deep insights into how the brain orchestrates complex psychological emotions and cognitive processes.

EEG signals are characterized by multiple frequency bands, among which Alpha and Beta waves are paramount, each corresponding to distinct functional states of the brain [34, 36, 62]. We categorize different EEG bands according to the following frequency definitions:

- Alpha Waves (8–13 Hz) are emblematic of the brain's state of relaxation and idleness. They are further categorized into three sub-bands based on their frequency range. 1) Alpha1 (8–8.9 Hz): This band is predominant when an individual is in a relaxed state with closed eyes, marking the onset of relaxation. 2) Alpha2 (9–10.9 Hz): These waves are more pronounced when the individual is relaxed yet maintains a level of alertness. 3) Alpha3 (11–12.9 Hz): This band appears as the brain relaxes further while remaining somewhat awake.
- Beta waves (13–30 Hz) are integral to the brain's alertness, focused attention, and cognitive processing. 1) Beta1 (13–18 Hz): Associated with mild cognitive activities and focused attention, like reading or simple thought processes. 2) Beta2 (18–22 Hz): These waves intensify during complex cognitive tasks such as problem-solving and decision-making. 3) Beta3 (22–30 Hz): This range signifies highly focused attention and rapid cognitive processing, indicative of active information processing.

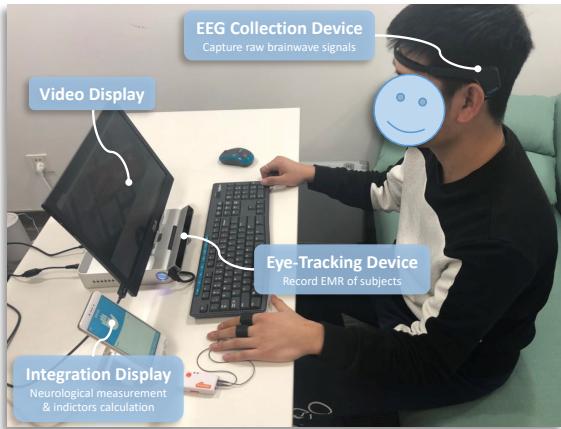


Figure 5. Equipment for Collecting Subjective Responses of SRI-ADV dataset. During data acquisition, participants wear an EEG Device, facing a Video Display, with an Eye-Tracking Device below to monitor gaze. Video durations and subjective responses are recorded on an Integration Display for analysis.

Table 7. Categories & Distribution of Subjectivity Task

Task	Engagement	Emotion	EMR
Cls-1	non-cognitive [0, 1) prop. 55.0%	negative (-∞, -6) prop. 29.7%	not attended [0, 0.45) prop. 29.3%
Cls-2	cognitive [1, +∞) prop. 45.0%	neutral (-6, 6) prop. 39.0%	partially attended [0.45, 0.6) prop. 41.6%
Cls-3	-	positive (6, +∞) prop. 31.2%	fully attended [0.6, +∞) prop. 29.0%
Distri.			

In this research, we analyze EEG signals and eye movement metrics to capture authentic subjective feedback from diverse demographic groups in response to advertisements. The SRI-ADV dataset, derived from raw signals and specific frequency bands, aids in improving the understanding of videos, especially those related to advertising.

B. Data distribution of Subjectivity Task in SRI-ADV

We initiate our analysis by performing a comprehensive histogram distribution assessment of Engagement, Emotion, and EMR indicators, anchored to Audience Profiles benchmarks. As delineated in Table 7, the concluding row graphically encapsulates the distribution of each evaluated SRI. The table's initial three rows delineate each category's designation, value range, and the proportion of data attributed to the corresponding category. For instance, the Engagement indicator is bifurcated into two categories: the first, termed "non-cognitive," spans a value range from 0 to 1 (non-inclusive), with 55.0% of observations classified under this category.

Leveraging the statistical insights derived, we proceed to discretize the extant SRI values into distinct categories, thereby facilitating the structuring of the Subjectivity Task associated with SRI-ADV.

C. Method Algorithm and Training Hyperparameters

We introduce the Hypergraph Multimodal Large Language Model (HMLLM), a novel approach designed to integrate and process multi-modal data, taking into full account both subjective and objective elements to comprehend advertising videos. Our training procedure is detailed in Algorithm 1. The method is grounded in the utilization of hypergraphs and large language model to effectively handle complex relationships within and across modalities.

We divide the training process into the following stages and adapt appropriate training hyperparameters, as shown in Table 8.

Algorithm 1: HMLLM: Hypergraph Multi-modal Large Language Model

```

Input : Video Key Frames  $F = \{f_0, f_1, \dots, f_N\}$ ,
          Textual Prompts  $T$ ,
          Ground Truth  $\mathbf{Y}_{gt}$ ,
          Warm Up Epoch  $E_0$ ,
          Fine-tune Epoch  $E_1$ ,
Initialization:  $\mathbf{X}_0 \leftarrow \text{Pre\_process}(F_v)$ 
                   $\mathbf{F}_v \leftarrow \text{Visual\_Encoder}(\mathbf{X}_0)$ 
                   $\text{SALM} \leftarrow \text{Initialize\_SALM}(\mathbf{F}_v, T)$ 
                   $\text{HL-Gate} \leftarrow \text{OFF}$ 
                   $\mathbf{Q} \leftarrow \text{Initialize\_Query}()$ 
// Stage I: SALM Warm Up
1 for  $i \leftarrow 1$  to  $E_0$  do
2    $\mathbf{K}, \mathbf{V} \leftarrow \text{QFormer}(\mathbf{F}_v)$ 
3    $\mathbf{F}_p \leftarrow \text{SALM\_Projector}(\mathbf{F}_v, \mathbf{K}, \mathbf{V})$ 
4    $\text{SALM} \leftarrow \text{SALM\_Train}(\text{SALM}, \mathbf{F}_p, T)$ 
5    $\tilde{\mathcal{Y}}_{qa} \leftarrow \text{SALM}(\mathbf{F}_p, T)$ 
6    $\mathcal{L}_{ITG} \leftarrow \text{ITG\_Loss}(\tilde{\mathcal{Y}}_{qa}, \mathbf{Y}_{gt})$ 
7    $\text{SALM\_Optimizer}(\mathcal{L}_{ITG})$ 
// Stage II: SAL-HL Fine-tuning
8  $\text{HL-Gate} \leftarrow \text{SRI\_Contained}(T)$  // Set
   HL-Gate ON.
9 for  $i \leftarrow 1$  to  $E_1$  do
10   $\mathbf{F}_{pv} \leftarrow \text{Feature\_Mixer}(\mathbf{F}_p, \mathbf{F}_v)$ 
11   $\mathbf{F}_{frame\_level} \leftarrow \text{Pool}(\mathbf{F}_{pv})$ 
12   $\mathcal{G} \leftarrow \text{Construct\_Hypergraph}(\mathbf{F}_{frame\_level}, \mathfrak{R})$ 
13   $\tilde{\mathcal{Y}}_{sri} \leftarrow \text{HGNN}(\mathcal{G}, \mathbf{F}_{frame\_level})$ 
14   $\mathcal{L}_{CE} \leftarrow \text{Cross\_Entropy}(\tilde{\mathcal{Y}}_{sri}, \mathcal{Y}_{sri})$ 
   //  $\mathcal{L}_{ITG}$  is obtained same with
   the stage I
15   $\mathcal{L} \leftarrow \text{Combined\_Loss}(\mathcal{L}_{ITG}, \mathcal{L}_{CE}, \lambda)$ 
16   $\text{Joint\_Optimizer}(\text{SALM}, \text{HGNN}, \mathcal{L})$ 

```

C.1. Initialization for Model Parameters

The process begins with the extraction of key frames from the input video, which are then pre-processed to standardize the input format. These initial visual features are encoded using a visual encoder, producing a set of feature vectors. Simultaneously, textual prompts are prepared for processing. A query set is initialized, marking the starting point for our model’s learning process.

C.2. Stage I: SALM Warm Up

Stage I is dedicated to the training of the SRI-Aware Language Model (SALM), with the goal of enhancing its capabilities in language generation and reasoning inference. In this stage, visual features are converted into key-value pairs using a query-former mechanism, essential for the attention processes. These features are then fed through the SALM projector, which enriches the model’s understanding by integrating textual prompts.

The training of SALM spans 10 epochs, as detailed in Table 8, with a particular focus on minimizing the Image-Text Grounding (ITG) loss. This step is crucial for ensuring that the model’s outputs are in alignment with the ground truth, thereby optimizing performance.

C.3. Stage II: SAL-HL Fine-tuning

Stage II shifts the focus to fine-tuning the SRI-Aware Language Hypergraph Learning (SAL-HL) component, with the objective of enhancing the model’s capability to mimic the subjective perceptual capacities of the brain. After completing the initial warm-up phase, the model enters the fine-tuning stage, marked by the activation of the Hypergraph Learning (HL) gate. This process enriches the model’s multi-modal context by combining these features into a pooled frame-level representation. This representation then forms the foundation for constructing a hypergraph that captures the intricate interconnections among data points.

Following the construction of the hypergraph, a Hypergraph Neural Network (HGNN) is employed to process the hypergraph structure, allowing the model to leverage the complex connections present within the data. The output generated by the HGNN is fine-tuned using Cross-Entropy loss along with the Image-Text Grounding (ITG) loss emphasized during the warm-up phase. This amalgamation of loss functions serves as a directive for the optimization process, targeting both the SRI-Aware Language Model (SALM) and the HGNN components. This strategic approach ensures a unified and coherent learning experience throughout the two distinct stages of the model’s training, fostering a comprehensive understanding and adaptation to the intricacies of the data.

Following [40], we incorporate Low-Rank Adaptation (LoRA) [29] modules into the SALM with a configuration of rank 16, an alpha value of 32, and a dropout rate of 0.1 during Stage II. Within the HGNN, it is imperative to adjust the input based on pooled frame-level representation, setting the Number of Vertices and Hyperedges to 8×98 and the Channel of Vertex Representation to 1024. For the hypergraph’s internal configuration, we adhere to the commonly used settings as illustrated in the Table 8, ensuring a balance between training effectiveness and model size.

Table 8. Training Hyperparameters for different stages.

config	Stage1 SALM Warm Up	Stage2 SAL-HL Fine-tune
input frame		8
input resolution		224
max text length		512
optimizer	AdamW	
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
weight decay	0.02	
learning rate schedule	cosine decay	
learning rate	1e-4	2e-5
batch size	128	64
warmup epochs	0.5	1
total epochs	10	20
λ of \mathcal{L}	0	0.1
augmentation	flip, MultiScaleCrop [0.5, 1]	
Vertices number of Hypergraph	-	8*98
Hyperedges number of Hypergraph	-	8*98
Channel of Vertex representation	-	1024
K of hypergraph construction	-	3, 4, 5
Average out-degree of vertices	-	12.5
In-degree of hyperedges	-	3, 4, 5

C.4. Summary

HMLLM stands as a comprehensive framework that leverages the strengths of hypergraph structures and multi-modal data integration. Through its two-stage training process, it achieves a deep understanding of the relationships within and across modalities, paving the way for advanced applications in multi-modal data processing and generation.

D. Computational Complexity and Resource Utilization of HMLLM

We have recorded the training time on eight A100 GPUs, each with 40GB of memory, and the inference time on a single A100 GPU with the same specifications. Our proposed HMLLM is generally on par with other models supporting video multi-modalities in terms of parameter count, training time, and inference time.

Table 9. Comparative Analysis of Model Parameters

Model	Video-Chat2	HMLLM (Ours)
Total Parameters (Billions)	7.2	7.2
Training Time (H/Epoch)	13.5	14.3
Inference Time (s/Video)	6.2	6.2
ACC on Task2	49.27	50.52

E. Zero-shot Prompt for Subjectivity Task

For the subjectivity task, we conducted Zero-shot inference on commercial and open-source models. During this pro-

cess, we tested various Prompts ^{2 3} to enable reasoning by the language models. The Prompts we selected are as follows.

As an AI model simulating EEG indicator analysis, your task is to systematically evaluate a user's feelings after viewing video or frames content using the provided EEG indicators: Cognitive Engagement (CE), Emotional Recognition (ER), and Eye Movement Ratio (EMR).

Effective analysis requires detailed video content information themes, narrative structure, visual and auditory elements and viewer attributes, including age, gender, preferences, and experiences. You will integrate the EEG indicators' definitions and probabilities to deduce the viewer's cognitive engagement, emotional response, and attention level.

This involves assessing how video elements may attract or repel the viewer, grounded in psychological principles and media consumption research. But you don't need to output the reasoning process, only the final result.

1. Cognitive Engagement (CE) Definitions:

- A. Non-Cognition:(55% probability)The viewer shows minimal interest or understanding, resulting in low EEG activity.

- B. Cognition:(45% probability)The viewer understands and relates to the video, evidenced by increased EEG activity.

2. Emotional Recognition (ER) Definitions:

- A. Negative:(30% probability)Dislikes certain video elements (e.g., conflict, unappealing objects).

- B. Neutral:(40% probability)Feels indifferent towards the video content (e.g., mundane tasks).

- C. Positive:(30% probability)Experiences enjoyment or excitement (e.g., appealing scenes).

3. Eye Movement Ratio (EMR) Definitions:

- A. Not Attended:(30% probability)Viewing ratio ≤ 0.45 , possibly due to unattractive visuals or cognitive dissonance.

- B. Partially Attended:(40% probability)Viewing ratio between 0.45 and 0.6, suggesting some attractive elements.

- C. Fully Attended:(30% probability)Viewing ratio > 0.6 , indicating high appeal and mood enhancement.

Please ensure your analysis follows this format with no additional output:

CE: B;

ER: B;

EMR: C

²<https://clickup.com/blog/ai-prompt-templates/>

³<https://www.zdnet.com/article/how-to-write-better-chatgpt-prompts-in-5-steps/>



Question: What does this video describe overall?

This video usually appears in the context of family and warm atmosphere, and may describe a warm scene of a family during winter, especially during **new years or holidays**. The video may showcase emotional communication among family members, such as the interaction between children and parents, grandmother and granddaughter, and the happy moments of family gatherings. The presentation of the product (possibly **milk or dairy products**) indicates that the video may be a **commercial advertisement** aimed at conveying the connection between the product and family health and happiness. The tail of the video may display the product brand and New Year's greetings, giving people a warm and peaceful feeling. The entire video may emphasize the warmth of family reunion and the joy of holiday moments.

Question: Please assess the engagement indicator after watching the video.

Engagement: cognitive

Question: If you are a male aged 36 to 40, how would your EMR duration change after watching this video?

EMR duration: fully attended

Question: After watching the video, what sort of trend may emerge in a 31 to 35-year-old woman's EMR duration?

EMR duration: partially attended



Question: What is the main audience for this video advertisement?

This video seems to be an **advertisement** that uses a comparison between a motorcycle rider and a desert camel, as well as some dramatic filming techniques, to attract the audience's attention. The motorcycle rider in the video seems to be exploring and traveling in a harsh desert environment. Finally, the advertisement showcases the product - a bottle of **functional beverage**, combined with slogans and visual elements - to convey that the product can enhance the endurance and physical function of motorcycle riders in extreme challenging environments, supporting users to achieve better results. Overall, this advertisement aims to showcase the advantages and characteristics of its product by highlighting themes of **adventure and endurance**.

Question: In terms of 31 to 35-year-old women, what is the typical performance they exhibit in the area of emotion post-video viewing?

emotion: neutral

A). Ground Truth

Gemini

This video shows a young boy looking out his window at the snow falling. He is smiling and appears to be happy. The video is set in a rural area, and the boy is wearing a traditional Chinese outfit. The video is likely intended to **evoke a sense of nostalgia and happiness**.

This video showcases a family enjoying a **snowy** day together, with scenes of a boy drinking **milk**, a woman drinking from a carton, and a man and woman drinking from a glass. The video also features a family enjoying a meal together at a table.

This video depicts a **celebration** of a holiday.

This video looks like a **commercial advertisement**. The video may showcase different scenes of a family enjoying **milk**, emphasizing the importance of milk in family life and the joy of family gatherings. The video may also include scenes of family members celebrating holidays or special occasions, such as the **New Year**, and may include product displays and brand information. Overall, the video aims to convey the **warmth** of family, the joy of holiday celebrations, and the role of milk in enhancing these beautiful moments.

Gemini

Engagement: cognitive

Engagement: non-cognitive

Engagement: cognitive

Engagement: cognitive

Gemini

EMR duration: partially attended

EMR duration: partially attended

EMR duration: fully attended

EMR duration: fully attended

Gemini

EMR duration: partially attended

EMR duration: partially attended

EMR duration: partially attended

EMR duration: partially attended



Gemini

Based on the video's dimensions, the main audience for this video advertisement is likely to be people who are **interested in traveling to or learning about the Gobi Desert**. The video is shot in a documentary style, and it provides a lot of information about the camels and the people who live in the desert. The video is also very visually appealing, and it is likely to make viewers want to learn more about the Gobi Desert.

The main audience for this video advertisement is likely people who enjoy adventure and exploration, as well as those who appreciate the beauty of nature and the thrill of **traveling on a motorcycle**. The video showcases the excitement and freedom of exploring the world on a motorcycle, and the adrenaline rush of traveling through different terrains.

The primary target audience for this advertisement is the Chinese audience.

This video advertisement is mainly aimed at outdoor adventure enthusiasts. The video shows a motorcycle rider and a camel, as well as a close-up of a bottle of product, which may be a **beverage** or **energy drink**. The video may emphasize the **vitality and adventure** of the product, and may be promoting a certain energy drink or other beverage that can provide energy or is related to motorcycle riding. The video may also emphasize the connection between the product and the lifestyle of motorcycle riders, aiming to attract this target audience.

Gemini

emotion: neutral

emotion: neutral

emotion: positive

emotion: neutral

Gemini Gemini-Pro-vision

Video-LLava

VideoChat2

HMLLM(ours)

B). Inference by Different Models

Figure 6. More qualitative analysis of SRI-ADV. Green signifies accurate descriptions, while red denotes incorrect or hallucinatory responses.

Finally, analyze the specified question without ex-

traneous , focusing on the indicators' specific index

based on probabilities.

{[The Question in Test Set]}

It is worth noting that to prevent the model from overfitting to specific choices, we randomized the options and tested them three times, taking the average result as the conclusive outcome.

F. Visualization and Qualitative Analysis

In the realm of advertising, the use of metaphors, scenic portrayals, and related content is prevalent. Our SRI-ADV dataset is meticulously crafted to support both subjective and objective analyses, thereby offering a comprehensive understanding of video advertising content. It uniquely bridges the gap between these analyses, with objective comprehension bolstering subjective interpretation. This fusion enables the exploration of qualitative aspects such as Engagement, Emotion, and Eye Movement Ratio (EMR) across various demographics.

As shown in Figure 6, Part A showcases the SRI-ADV’s ground truth, distinguished by its detailed annotations and extensive response lengths. Meanwhile, Part B delineates a comparative analysis among the outputs generated by the Gemini-Pro-vision, Video-LLaVA, and VideoChat2 models against our HMLLM.

For instance, an energy drink advertisement as shown in the bottom of Figure 6, HMLLM uniquely captures both the overt (a motorcycle rider and a camel) and the covert (the product’s essence of vitality and adventure) elements of the advertisement. This comprehensive analysis extends to the advertisement’s main audience, design principles, visual narratives, and product attributes, showcasing our model’s superior capability in extracting and interpreting complex thematic elements.

A Comprehensive Survey on EEG-Based Emotion Recognition: A Graph-Based Perspective

Chenyu Liu*

College of Computing and Data
Science, Nanyang Technological
University
chenyu003@e.ntu.edu.sg

Yi Ding

College of Computing and Data
Science, Nanyang Technological
University
ding.yi@ntu.edu.sg

Xinliang Zhou*

College of Computing and Data
Science, Nanyang Technological
University
xinliang001@e.ntu.edu.sg

Yihao Wu

College of Computing and Data
Science, Nanyang Technological
University
yihao005@e.ntu.edu.sg

Liming Zhai

School of Computer Science, Central
China Normal University
limingzhai@ccnu.edu.cn

Kun Wang†

University of Science and Technology
of China
wk520529@mail.ustc.edu.cn

Ziyu Jia†

Institute of Automation, Chinese
Academy of Sciences
jia.ziyu@outlook.com

Yang Liu

College of Computing and Data
Science, Nanyang Technological
University
yangliu@ntu.edu.sg

Abstract

Compared to other modalities, electroencephalogram (EEG) based emotion recognition can intuitively respond to emotional patterns in the human brain and, therefore, has become one of the most focused tasks in affective computing. The nature of emotions is a physiological and psychological state change in response to brain region connectivity, making emotion recognition focus more on the dependency between brain regions instead of specific brain regions. A significant trend is the application of graphs to encapsulate such dependency as dynamic functional connections between nodes across temporal and spatial dimensions. Concurrently, the neuroscientific underpinnings behind this dependency endow the application of graphs in this field with a distinctive significance. However, there is neither a comprehensive review nor a tutorial for constructing emotion-relevant graphs in EEG-based emotion recognition. In this paper, we present a comprehensive survey of these studies, delivering a systematic review of graph-related methods in this field from a methodological perspective. We propose a unified framework for graph applications in this field and categorize these methods on this basis. Finally, based on previous studies, we also present several open challenges and future directions in this field.

CCS Concepts

- Computing methodologies → Artificial intelligence; • Human-centered computing → Human computer interaction (HCI).

Keywords

EEG, Emotion Recognition, Graph

*Co-author.

†Corresponding author.

1 Introduction

Emotion is an integral and complex aspect of human cognition, playing a crucial role in decision-making, behavior, and social interactions [1]. Consequently, emotion recognition is essential for mental health diagnosis and human-computer interaction [122]. In this context, the direct correlation between electroencephalogram (EEG) signals and brain activity has established EEG-based emotion recognition as a highly specific and valuable task. First, EEG is a manifestation rather than an expression of emotion, which is more likely to reveal the genuine emotional state than other modalities objectively. The expression of emotion includes facial micro-gestures [102], embodied behavior [13], gesture [40], speech intonation and voice quality [24], which serve a communicative function and are largely under conscious deliberate control [31]. In addition, EEG recordings provide direct measures of neural activity, offering a more accurate representation of brain emotional states than other physiological manifestations of emotion. Physiological metrics such as skin conductivity, temperature, heart rate, etc., are manifestations of physiological systems in emotional states that do not reflect emotional activity in the human brain directly. Therefore, EEG-based emotion recognition tasks hold irreplaceable significance in the study of genuine human emotions and the exploration of emotional activities within the human brain.

In the field of EEG-based research, emotion recognition requires more attention to dependency between brain regions compared to other paradigms [41]. As illustrated in Figure 1 (a), the human brain functions as a complex network with hierarchical and functional organization at the level of brain regions [84]. These regions are responsible for processing specific functions and are particularly relevant to various EEG-based paradigms, as shown in Figure 1 (b). Specifically, the speech decoding paradigm focuses on the temporal lobe, which is responsible for language comprehension and processing [26]. The primary motor cortex, located in the frontal lobe,

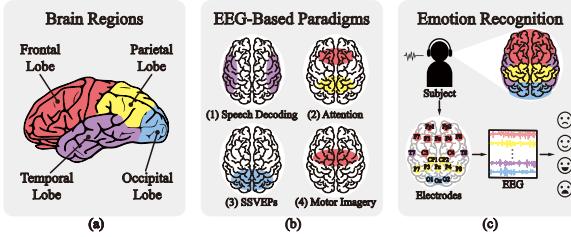


Figure 1: Description of brain regions and EEG-based paradigms. (a) Four regions of the cortex. (b) Four other common EEG-based paradigms. (c) The process of EEG-based emotion recognition.

regulates motor functions and is thus relevant to motor imagery paradigms [45]. Visual stimulation activates the occipital lobe, the primary brain region of interest for the steady-state visually evoked potentials (SSVEP) paradigm [91]. However, as shown in Figure 1 (c), emotions are high-level neurocognitive functions expressed as cognitive-emotional interactions generated from brain regions with a high degree of connectivity [82]. These interconnected brain regions do not function independently [21]. Activating a particular brain region often leads to the activation of other regions within the connected network [42]. Thus, developing neurophysiologically meaningful networks to effectively model the connectivity among functional brain regions during emotional states is central to the field.

A significant trend in modeling the aforementioned connectivity is the application of graphs. Graphs excel at representing relational data through nodes and edges, making them particularly suitable for modeling the complex interactions between brain regions during emotional states. The local or distributed processes between brain regions during emotional states [49] can be effectively modeled by edge-based node aggregation [41], where the edge represents the connectivity between brain regions. As a result, an increasing number of methods have employed graphs to capture the intricate connectivity patterns inherent in emotion-related EEG, which are crucial for enhancing the accuracy of emotion recognition [27]. These approaches are specifically developed from diverse perspectives to integrate brain physiology paradigms, distinguishing them structurally from graph-based techniques in other fields. Nonetheless, there is currently no standardized framework for the application of graphs in EEG-based emotion recognition. This lack of standardization highlights the need for a comprehensive survey that encompasses various graph-based methodologies in this field.

To this end, this paper presents a systematic survey of graph applications in EEG-based emotion recognition. Our aim is to offer comprehensive guidance on constructing graph-based models in this field. The contributions of this survey are summarized as follows:

- **The first survey.** This survey provides a comprehensive and systematic review of graph-related methods in EEG-based emotion recognition. To the best of our knowledge, this is the first and only survey work on such a topic.

- **Systematic methodology-centric taxonomy.** This survey introduces a unified framework from a methodological standpoint on graph application and provides a clear guideline for the construction of graph-related methods in this field.
- **Future research directions.** This survey summarises and highlights future directions to facilitate graph application in this field.

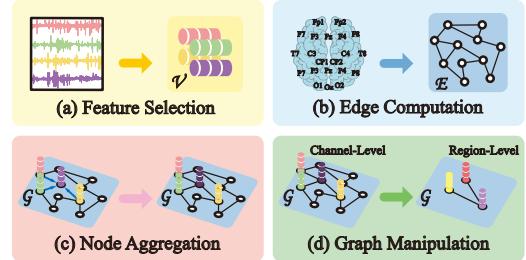


Figure 2: A Unified framework of graph-related methods in EEG-based emotion recognition.

2 Background

2.1 EEG-based Emotion Recognition

EEG-based emotion recognition leverages the close relationship between EEG signals and brain activity to accurately identify emotional states. As shown in Figure 1 (c), emotions are complex neurocognitive processes involving dynamic interactions among multiple brain regions, such as the frontal lobes, parietal lobes, and temporal lobes [86]. Each brain region's activity during emotional states contributes uniquely to the overall emotional experience, with specific areas playing distinct roles in the processing of different emotions [5, 32, 88]. These regions do not operate in isolation; rather, they interact in a highly coordinated manner, forming a network of activity that underlies the processing and regulation of emotions [83]. The various channels of an EEG can measure the neural activity of these corresponding regions, capturing real-time changes in brain dynamics [2]. Thus, emotion-related EEG can reflect the functional connectivity between different brain regions during emotional states, providing a comprehensive view of how emotions are processed in the brain.

2.2 Preliminaries

Graph-related methods in EEG-based emotion recognition can be indicated as taking EEG signal $\mathbf{X} \in \mathbb{R}^{C \times S}$ as input, where C is the number of channels and S is the number of samples, constructing graphs and predicting emotion labels $\mathbf{Y} \in \mathbb{R}^1$. We consider an emotion-related graph denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where \mathcal{V} and \mathcal{E} represent the sets of nodes and edges, respectively. The feature matrix of \mathcal{V} is denoted as $\mathbf{V} \in \mathbb{R}^{C \times D}$, where D is the dimension of node features. An adjacency matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$ represent the connections between nodes, where $a_{i,j} = 1$ if $(v_i, v_j) \in \mathcal{E}$. To learn the node representations in a graph \mathcal{G} , most methods adhere to the following paradigm of neighborhood aggregation and message passing:

$$\mathbf{h}_i^{(l)} = \text{COMB} \left(\mathbf{h}_i^{(l-1)}, \text{AGGR} \{ \mathbf{h}_j^{(l-1)} : v_j \in \mathcal{N}(v_i) \} \right), \quad 0 \leq l \leq L \quad (1)$$

where L is the number of network layers, $\mathbf{h}_i^{(l)}$ ($1 \leq l \leq L$) denotes the node embedding of v_i at the l -th layer. **AGGR** and **COMB** represent functions used for aggregating information from neighborhood nodes $\mathcal{N}(v_i)$ and combining ego- and neighbor-representations, respectively.

3 Taxonomy

Recent advancements in graph application have demonstrated their great potential to achieve better performance in EEG-based emotion recognition. These methods focus on different aspects of the graph but generally address a few key questions: (1) what kind of EEG features should be selected; (2) how to compute the adjacency matrix to characterize the connectivity of brain regions in emotional states; (3) how to perform node aggregation; and (4) what kind of graph manipulation should be adopted. Based on these questions, we summarized and categorized the existing studies based on a unified framework as shown in Figure 2.

The proposed taxonomy is illustrated in Figure 3, and the related works can be found in Table 1. The proposed taxonomy offers a structured and comprehensive classification to deepen the understanding of graph-based approaches in EEG-based emotion recognition. It is organized into four hierarchical levels, starting with the feature selection, followed by the edge computation, node aggregation, and finally, the graph manipulation. **(a) Feature selection** indicates the type of node features selected, which contains temporal and frequency nodes; **(b) Edge computation** specifically refers to calculating the adjacency matrix, which is further classified into model-dependent and model-independent edges according to the participation of parameters. **(c) Node aggregation** denotes the updating method of nodes, which is categorized into spectral-based and spatial-based methods. **(d) Graph manipulation** represents the manipulation of graph structures at the network level, including multi-graphs, hierarchical graphs and spatio-temporal graphs. In the following, we introduce the four stages and their subcategories in more detail.

4 Feature Selection

The unique characteristics of emotion-related EEG have led to the development of distinct methods for node feature selection in graph-based approaches within this field. As EEG is a time series data, the most intuitive approach is to utilize temporal features. Additionally, due to the direct correlation between EEG frequency and emotional activity [4], some methods employ unique approaches to extract frequency domain features as nodes. In this section, we classify the types of nodes employed in graph-based methods into temporal and frequency nodes. In the following, we provide a detailed introduction to these two types.

Temporal Nodes: This is the most intuitive and straightforward method, where nodes directly use raw signals or time-related features. For example, LGGNet [15] and SCC-MPGCN [118] input the raw EEG signals into the network after basic filtering. Raw EEG signals facilitate joint analysis with other physiological signals. HetEmotionNet [35] and VBH-GNN [67] incorporate temporal

physiological signals such as Electrocardiogram (ECG) and Galvanic Skin Response (GSR) as auxiliary modalities for synchronous time-domain analysis. Additionally, Sparse-DGCNN [115] employs Amplitude Spectrum Mean (ASM), using the average amplitude spectrum of EEG signals as node features.

Frequency Nodes: The fundamental reason for utilizing frequency features as nodes is that there exists a direct connection between different frequency bands of EEG and different emotions [4, 70], such as the β -band is associated with anxiety. Therefore, frequency nodes typically focus on five specific bands: delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-50 Hz) bands. The most commonly used frequency node at present is the Differential Entropy (DE) feature [17]:

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

where μ and σ denote the parameters of Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ that the EEG obeys. Many methods, such as [16, 64, 78], estimate the probability density function of the signal and use the aforementioned formula to calculate the DE for each band as node features. Furthermore, methods such as [60, 95, 116] utilize combinations of DE features from symmetrical electrodes (differential asymmetry and rational asymmetry). In addition, methods such as [37, 61, 73] use the Power Spectral Density (PSD) of EEG signals. A few methods employ other frequency domain features, such as the Differential Cumulative Average of Uniformity (DCAU) used by DGCNN [98] and the Short-Time Fourier Transform (STFT) used by GDDN [8].

Remark. Although the majority of current research employs frequency nodes, we argue this prevalence is not due to the superiority of frequency nodes over temporal nodes but is instead influenced by the datasets. For instance, methods using the SEED [119] dataset commonly adopt DE features, while those using the MPED [97] dataset typically use STFT. Only a few methods utilizing the DREAMER [93] and DEAP [43] datasets employ temporal nodes. The high usage rate of the SEED dataset has contributed to the prevalence of frequency nodes. It is evident that frequency nodes can mitigate the effect of transient noise (e.g., eye movements) across multiple frequency bands, thereby providing a more stable representation of emotional states. However, apart from their end-to-end implementation, temporal nodes exhibit greater potential in reflecting event-related potentials in EEG under emotional activity, such as transient changes in amplitude and waveform [90]. Additionally, temporal nodes offer higher extensibility, as they can accommodate other physiological signals for joint analysis [104].

5 Edge Computation

Edge computation poses a significant challenge for graph-based methods in EEG-based emotion recognition. On the one hand, the edges in emotion-related graphs have inherent neurophysiological foundations. For instance, the electrode positions of EEG acquisition devices generally conform to the 10-20 system, which dictates

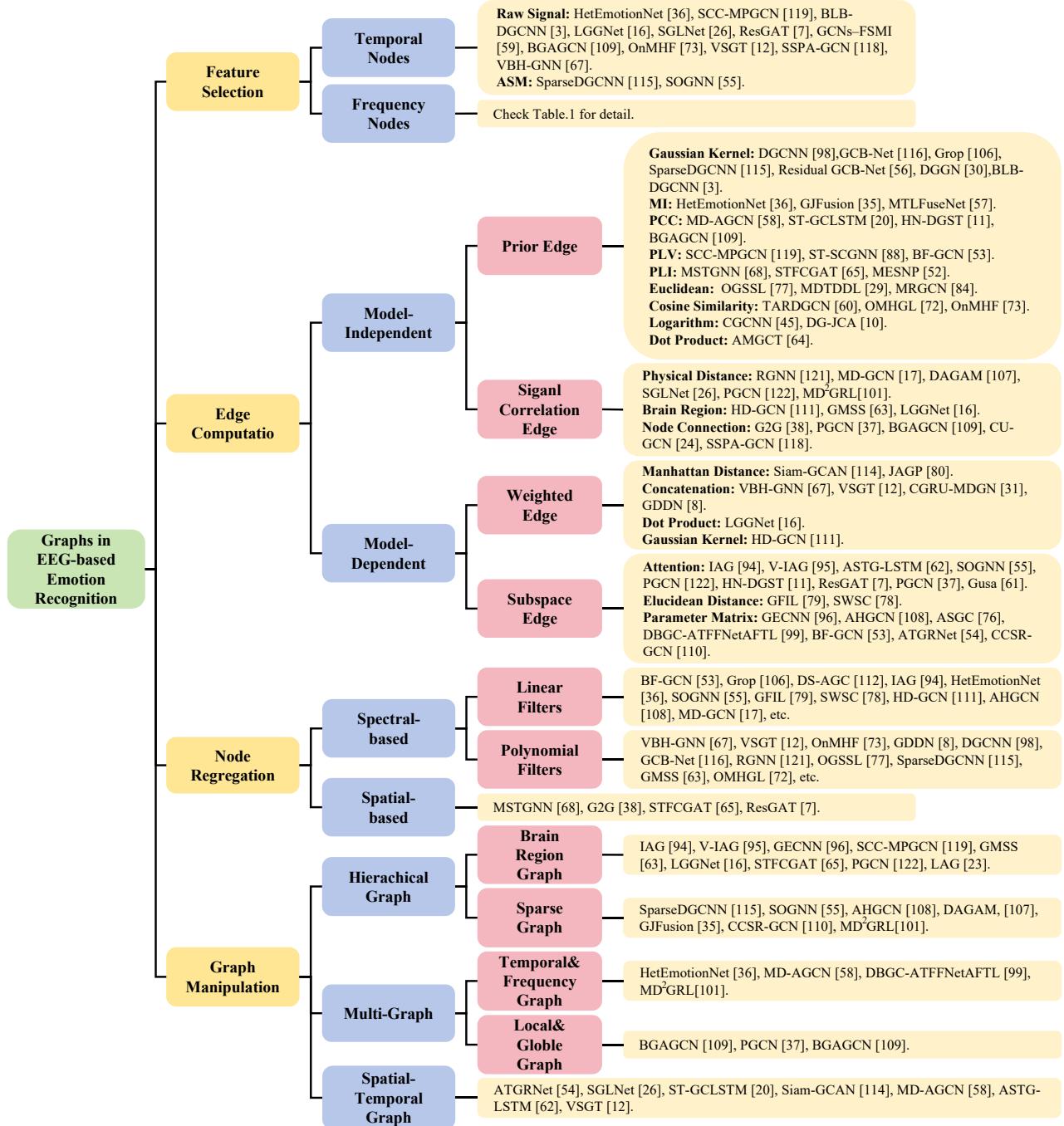


Figure 3: An overview of the categorization.

fixed distances and connectivity between nodes, each representing different brain regions. On the other hand, the dependencies between brain regions under emotional states are highly variable. Firstly, there are differences in brain region dependencies across different emotional states [82]. Secondly, even under the same emotional state, different subjects exhibit variations in the intensity and patterns of their emotional brain activities [65]. Additionally, for the

same subject, brain activity patterns can vary when experiencing the same emotion at different times due to differences in response intensity and psychophysiological states [100]. Therefore, graph-based methods in this field strive to balance the fixed and variable nature of edges. In this section, we categorize these methods into two types based on whether model parameters are involved in edge computation: model-independent and model-dependent methods.

In the following, we describe these two methods and their subtypes in detail.

5.1 Model-Independent Edge

The computation of model-independent edges does not involve model parameters. These edges aim to maximally preserve the neurophysiological foundations of emotion-related edges, thus having clear physiological structures and physical significance. The existing literature typically adopts the following two strategies.

Prior Edge: In this method, edges are manually defined based on electrode connections from physiological paradigms. An example of prior edges is in [16, 107, 120], which utilize the theory that the strength of connections between brain regions decays as a function of the inverse square of the physical distance [87]. These methods use 3-D coordinates to calculate the physical distances between electrodes as the edges. Methods such as [37, 109, 121] determine the connectivity between electrodes directly based on the electrode arrangement in the International 10-20 system, thereby constructing the adjacency matrix. Based on this, [15, 62, 111] restrict the connectivity between electrodes within specific groups, considering the distribution of brain regions and the symmetry between the left and right hemispheres.

Signal Correlation Edge: The computation in this method is based on signal processing, using the similarity or distance between EEG signals from different channels as edges. The propagation of physiological electrical signals is diffusive [76], which means that the signal collected by one electrode will contain EEG signals from adjacent electrodes. Therefore, signal correlation edges preserve the neurophysiological foundations between electrodes through signal-level similarity. [28, 78, 85] use the Euclidean distance between node signals as edges. Similarly, [59, 73] employ cosine distance, while MSTGNN [68] uses the Phase Lag Index (PLI). Additionally, [11, 19, 57, 109] utilize the Pearson Correlation Coefficient (PCC) to compute the strength of the linear relationship between signals as edges. [51, 52, 72, 118] use the Phase Locking Value (PLV) to calculate the degree of phase similarity between electrode signals at a particular frequency as their edges. Other methods include Gaussian kernel, such as [98, 115, 116], and Mutual Information (MI), as used by [34, 35, 56].

Remark. The limitation of model-independent edges is that, although they preserve the basic connectivity of electrodes under emotional states, it is challenging to model the synergistic relationships between brain regions during emotional activities. Even the relatively flexible signal correlation edges, which use data-driven computation methods to reflect differences between different individuals and emotional states, are still insufficient to capture the complex spatial-temporal dependencies. However, the relatively fixed learning approach of model-independent edges does not mean they cannot be optimized during training. A common update method is:

$$\mathbf{A}' = (1 - \rho)\mathbf{A} + \rho \frac{\partial \text{Loss}}{\partial \mathbf{A}}, \quad (3)$$

where ρ denotes the learning rate. This allows model-independent edges to adapt to the variability of EEG under emotional states to some extent and ensures stable performance in scenarios with limited training data.

5.2 Model-Dependent Edge

The introduction of parameters allows model-dependent edges to dynamically fit the complex dependencies in emotion-related EEG. Model-dependent edges can adjust and refine the relationships between nodes, which represent different brain regions, based on the varying intensities and patterns of emotional activities. The existing literature typically adopts the following two strategies.

Weighted Edge: In this method, model parameters are introduced as a weight matrix in the computation, transforming the connection between nodes into nonlinear ones through activation functions. It can be regarded as a special case of signal correlation edges combined with model parameters. A common computation method for weighted edges is:

$$e_{ij} = \frac{\exp(\sigma(\omega^T \mathcal{R}\{x_i, x_j\}))}{\sum_{j=1}^C \exp(\sigma(\omega^T \mathcal{R}\{x_i, x_j\}))} \quad (4)$$

where $\sigma(\cdot)$ represents an activation function, such as ReLU. x_i and x_j represent the node embeddings of v_i and v_j , and ω denotes the model parameters. $\mathcal{R}\{\cdot, \cdot\}$ represents a specific operation on the node embedding pairs. For example, in [81, 114], it represents the computation of the Manhattan distance, while in [71, 112], it represents the computation of the Euclidean distance. Additionally, some methods, such as [8, 30], concatenate the node embeddings directly for linear transformation, with VBH-GNN concatenating the node embeddings along the channel dimension and using convolutional layers instead of linear transformations to reduce the number of parameters.

Subspace Edge: This method uses model parameters to project the node embeddings into a subspace, where the dot product of the projected nodes is used as edges. The process can be described as:

$$e_{i,j} = \sigma((x_i \omega_i)(x_j \omega_j)^T), \quad (5)$$

where ω_i and ω_j represent the projection matrices for two nodes, respectively. For example, [94, 95, 121] adopt the above computation formula. Fundamentally, the dot product in the projection space is consistent with the dot product of the Query and Key matrices in the attention mechanism. Therefore, some methods, such as [7, 11, 54, 61], directly use the output of the attention mechanism as edges.

Remark. Model parameters allow these edges to dynamically fit the complex spatial-temporal dependencies in emotion-related EEG. For example, as a special case of signal correlation edges, weighted edges create more flexible nonlinear connections. However, the introduction of more model parameters increases the risk of overfitting. This is particularly true for subspace edges, which lack specific node relationship constraints, making it difficult to generalize in scenarios with insufficient training data. Additionally, some methods, such as [53, 77, 108, 110], randomly initialize a parameter matrix as the adjacency matrix. This entirely data-independent approach is even more challenging to train. Although many previous studies have not opted for model-dependent edges in subject-dependent scenarios, we argue that this parameterized, data-driven edge computation method is the future trend for EEG-based emotion recognition. For example, for cross-dataset or cross-subject

domain adaptation scenarios or emotion foundation models, adopting model-dependent edges can more accurately capture the underlying dependencies in emotion-related EEG data. Moreover, one potential approach is to integrate more neurophysiological foundations into model-dependent edges. For instance, VSGT [66] simulates the superposition state of neuronal electrical signals using Gaussian approximation and introduces parameters through re-parameterization [38] to update model-dependent edges.

6 Node Aggregation

Graph-based methods in this field commonly utilize convolutional operations to aggregate node features. These methods generate representations for node v_i by aggregating its own features h_i and the features of its neighbors h_j where $v_j \in \mathcal{N}(v_i)$. Consequently, these methods employ a fixed number of convolutional graph layers to extract high-level node representations, thereby capturing the dependencies between brain regions under emotional states. We categorize these methods into spectral-based and spatial-based approaches. In the following sections, we delve into the specifics of these two categories and their subtypes.

6.1 Spectral-based Method

Spectral-based methods capture the dependencies between brain regions under emotional states by transforming all node features into a weighted sum of different feature vectors through the graph Fourier transform. This means that the features of all EEG channels are filtered through a fully connected feature space, with messages being simultaneously transmitted between all nodes, thereby simulating the global coordination between brain regions under emotional states. We define the convolution process of this aggregation method as:

$$h_i^{(l)} = \sigma \left(\mathbf{U} \sum_{j=1}^C \mathbf{G}_{ij}^{(l-1)} \mathbf{U}^\top h_j^{(l-1)} \right), \quad (6)$$

where σ is the activation function, $h_i^{(l-1)}$ denotes the node embedding of v_i at the l -th layer, and $\mathbf{G}_{ij}^{(l-1)}$ represents a diagonal matrix with learnable parameters. This process can also be seen as using \mathbf{G} as a filter to remove noise from the emotion-related EEG graph signals. The existing research employs the following two types of filters.

Polynomial Filters: In this method, the filter is replaced by a simplified polynomial function, avoiding explicit eigenvalue decomposition and thereby reducing the time complexity. In this field, all methods using polynomial filters, such as [53, 95, 98], employ Chebyshev expansion mentioned in ChebNet [14] to construct the filter:

$$h_i^{(l)} = \sigma \left(\mathbf{U} \left(\sum_{k=1}^K \theta_k \mathcal{T}_k(\tilde{\Lambda}) \right) \mathbf{U}^\top h_j^{(l-1)} \right) \quad (7)$$

where $\tilde{\Lambda} = \frac{2\Lambda}{\lambda_{\max}} - \mathbf{I}_N$ is the eigenvalue matrix scaled in the range $[-1, 1]$. Λ is a diagonal matrix of eigenvalues. $\mathcal{T}_k(x) = 2x\mathcal{T}_{k-1}(x) - \mathcal{T}_{k-2}(x)$ constructs the orthogonal space, and $\mathcal{T}_k(\tilde{\Lambda})$ is the k -th order Chebyshev polynomial at $\tilde{\Lambda}$.

Linear Filters: Linear filters are a further simplification of polynomial filters. Methods that use linear filters, such as [72, 81, 112],

similar to GCN [39], adopt the first-order approximation of the Chebyshev expansion as the filter. Their convolution process can be summarized as follows:

$$h_i^{(l)} = \sigma(\bar{\mathbf{A}} h^{(l-1)} \omega_i), \quad (8)$$

where ω_i represents the weight matrix. $\bar{\mathbf{A}} = \mathbf{I}_N + \mathbf{D}^{-(1/2)} \mathbf{A} \mathbf{D}^{-(1/2)}$ is the renormalization trick of the adjacency matrix. Obviously, the graph filter is linear with the input adjacency matrix. From a spatial-based perspective, these methods can be considered as aggregating feature information from a node's neighborhood.

Remark. Currently, there has been limited innovation in the design of filters within graph-based methods in this field. Instead, researchers continue to use established methods such as ChebNet and GCN. However, even GCN, an improved form of ChebNet, has seen numerous advancements and modifications within the graph neural network domain due to its inherent limitations. The brain region activities under emotional states are complex and dynamic, making it challenging for the simple linear filters of GCN to effectively capture the relationships between EEG channels. This is because the predefined filters in GCN are typically designed for first-order information, which cannot directly capture such high-order relationships. In summary, there is a need for further exploration and improvement in the selection and design of filters, as well as in the development of methods specifically tailored for emotion-related EEG. Combining linear filters with complex filters to learn high-order brain region dependencies should be a key focus for EEG-based emotion recognition.

6.2 Spatial-based Methods

Spatial-based methods define node aggregation based on the spatial relationships between nodes. Specifically, these methods convolve the node representation of the central channel with those of its neighboring channels to derive an updated representation for the central channel, thereby propagating information along the connectivity of the electrodes. The primary distinction between spatial-based and spectral-based methods lies in their respective modes of information propagation between nodes. Methods such as [7, 37, 64] incorporate attention mechanisms during the propagation process to assign weights to the contributions of neighboring nodes to the central node. Consistent with the GAT [103], their update process can be summarized as:

$$h_i^{(l)} = \sigma \left(\sum_{v_j \in \mathcal{N}(v_i)} \alpha_{ij} \omega_i h_j^{(l-1)} \right), \quad (9)$$

where $\alpha_{ij} = \sigma(\text{ATT}(h_i^{(l-1)}, h_j^{(l-1)}))$ represents the attention weight, or connective strength, between v_i and v_j . Similar to linear filters, these methods update node embeddings through linear transformations. However, their adjacency matrices, which denote the existence of edges between nodes, limit the update scope of the central node to its connected neighbors. This results in sparse adjacency matrices that emphasize local spatial dependencies.

Another method is grounded in the theory that there exist inherent connections and pathways in the brain during emotional activities [6]. MSTGNN [68] utilizes the minimum spanning tree to

simulate pathways between electrodes. The pathways are defined as the routes through which electrical signals propagate during emotional activities. Within a pathway, each node can have up to one parent node and multiple child nodes. Consequently, the update method is:

$$h_i^{(l)} = \sigma(\omega_i \left((1 + \varepsilon^{(l-1)}) \cdot h_i^{(l-1)} + \omega_f h_f^{(l-1)} + \omega_c \text{MEAN}(h_c^{(l-1)}) \right)), \quad (10)$$

where $h_f^{(l-1)}$ represents the node embedding of the parent node. If v_i is the root node, the parent node embedding is set to zero. $h_c^{(l-1)}$ represents the node embeddings of the child nodes $v_c \in N(v_i)$. MEAN denotes averaging over all child nodes. ω_f and ω_c are learnable parameters for parent and child nodes.

Remark. In EEG-based emotion recognition, the boundaries between spatial-based and spectral-based methods are becoming increasingly blurred. This is because node aggregation in spatial-based methods is essentially an approximate linear process, which aligns with the principles of linear filters used in spectral-based methods. Consequently, employing an attention mechanism to infer emotion-related EEG edges followed by node aggregation using spectral-based methods is fundamentally and practically similar to first determining electrode connectivity and then incorporating attention mechanisms within spatial-based aggregation.

A promising approach is to integrate neurophysiological foundations into spectral-based methods. For example, MSTGNN uses Minimum Spanning Trees (MST) to simulate brain pathways under emotional states. This aligns with the perspectives introduced in edge computation, emphasizing the incorporation of neurophysiological foundations into model parameters. By doing so, models can more accurately capture the intricate dependencies and interactions that characterize emotional brain activity, enhancing both the robustness and interpretability of EEG-based emotion recognition systems. In summary, the convergence of spatial-based and spectral-based methods, along with the integration of neurophysiological insights, holds significant potential for advancing the effectiveness of EEG-based emotion recognition models.

7 Graph Manipulation

In EEG-based emotion recognition, graph-related methods frequently implement specific model-level graph manipulations to enhance the model's ability to represent and learn the complex dependencies inherent in the data. These manipulations address the multifaceted nature of emotion-related EEG, which includes variations across both the frequency and temporal domains, as well as intra- and inter-regional brain connections. Such structural adjustments are crucial for tailoring graph representations to better align with the neurophysiological foundations of emotional brain activity. Existing literature frequently adopts the following three strategies:

7.1 Hierarchical Graph

Hierarchical Graphs leverage the natural grouping of nodes to align with the physiological foundations of emotion-related EEG, particularly the distribution of brain regions. This approach aims

to construct the spatial dependencies of emotion-related EEG from a more macroscopic perspective. Based on the different types of dependencies established by this method, we categorize existing research into two types.

Brain Region Graph: In this method, channel-level nodes are grouped to further infer inter-regional dependencies at the brain region level. These dependencies reflect the coordination between different brain areas, such as the interactions between the frontal cortex and the amygdala during emotion regulation [88]. One type of method employs a manually established hierarchy based on the distribution of electrodes corresponding to different brain regions. For example, [94, 95] infer fully connected spatial dependencies in emotion-related EEG and then divide the brain regions into 17 groups according to the 10-20 system, averaging the embeddings of channels within each group to obtain regional embeddings. Another strategy, adopted by methods such as [22, 62, 64, 96], is to first group nodes based on brain region distribution to learn intra-regional spatial dependencies and then use the regional graph embeddings as nodes in a global graph to learn inter-regional relationships. In addition to grouping based on brain regions, some methods like [15, 121] also divide the hierarchy according to the left and right hemispheres of the brain. The rest method involves adaptively grouping channels and then inferring spatial dependencies between these groups. For instance, [108, 118] set a parameter weight matrix to linearly transform the channel-level node embeddings, reducing the dimensionality of the graph embeddings from the number of channels to the number of brain regions to achieve hierarchical division.

Sparse Graph: This method aims to represent the small-world topology [75] of brain regions during emotional activities, characterized by densely clustered local connections with sparse long-range connections [20]. The underlying theory is that certain brain functions activate only a limited number of brain regions. Thus, sparse cortical activity can explain the EEG patterns generated by deeper sources [47]. For example, MD²GRL [101] uses a parameter as a threshold to filter and retain a specific number of nodes, thereby achieving sparse connectivity among brain regions. Similarly, methods in [107, 118] learn a variable weight for all nodes, selecting a specific number of nodes based on ranking. Additionally, conventional pooling methods can achieve similar results, as demonstrated by GJFusion [34].

Remark. These two types of methods represent efforts in the field to combine neurological foundations with network structures. We argue that integrating them could potentially offer significant benefits. Currently, the approach of retaining a specific number of nodes in sparse graphs inevitably leads to the loss of potentially valuable information. Therefore, basing the model on the brain region graph and then inferring sparsity on the coarsened graph could maximize the retention of channel information while simulating the sparse connectivity of brain regions.

7.2 Multi-Graph

Multi-Graph allows models to capture different emotion-related EEG dependencies simultaneously by concatenating multiple types of graph embeddings. This method leverages the complementary

information provided by different graph structures, enriching the feature space and improving the model’s ability to discern emotional states. Based on the complementary domains, we further introduce two types of this method in the following.

Temporal&Frequency Graph: In this method, the model employs parallel graph structures to build the dependencies of emotion-related EEG in both the temporal and frequency domains. Temporal dependencies enable the model to detect amplitude features that are closely correlated with brain region, while frequency dependencies capture the activation of specific frequency bands associated with different emotional states, such as increased energy in the alpha band during sadness [18, 92]. Methods in [35, 101] integrate temporal&frequency graphs at the graph embedding level. They utilize a two-stream structure, where each stream corresponds to the temporal-spatial and frequency-spatial domains, respectively, and compute its own adjacency matrix. The graph embeddings from both streams are then concatenated and fed into a classifier. A distinct strategy employed by [57, 99] involves merging the adjacency matrices of the two streams before the node aggregation, resulting in a shared hybrid adjacency matrix.

Local&Globale Graph: This graph is less commonly applied in this field. Since it distinguishes itself from brain region graph methods only by decoupling the inference of inter-regional and intra-regional dependencies into two parallel processes. In the Local stream, electrodes are grouped based on brain regions, and connectivity is confined within these regions. Methods such as [36, 109, 111] utilize this strategy to focus on the local dependencies within specific brain areas. MRGCN [85] replaces the fixed brain region divisions with a more flexible framework. It introduces short-range and long-range spatial dependencies, which correspond to localized intra-region correlations and inter-region correlations, respectively.

Remark. Currently, multi-graph methods in the field struggle to ensure that the accuracy improvements gained from multi-stream structures justify the additional computational burden. This challenge arises because most multi-graph implementations simply merge two types of graph embeddings without fully optimizing the interaction between them. However, it is undeniable that multi-graph methods effectively leverage the complementary nature of different dependencies in emotion-related EEG data.

7.3 Spatial-Temporal graph

In this method, emotion-related EEG is decomposed into multiple time slices to construct temporal dependencies between these slices. This method views emotions as dynamic processes rather than static states. By using a temporal encoder, it captures the changes in these spatial dependencies over time to infer emotion labels. The difference between the current methods lies mainly in the selection of the encoder. For example, [19, 25, 61] employ LSTM as the encoder for sequential data. In HetEmotionNet, the LSTM is replaced with Gated Recurrent Units, while ATGRNet [53] uses a Temporal Convolutional Network (TCN). Some relatively simplified approaches, such as Siam-GCAN [114], concatenate all the graph embeddings and feed them into a fully connected layer. MD-AGCN [57] averages the graph embeddings of all time slices for further

classification.

Remark. Although the use of spatial-temporal graphs is currently limited, we argue this approach holds significant potential in this field. Emotions are dynamic processes that evolve with the development, intensification, and subsidence of internal and external stimuli [33, 46]. Therefore, this method aligns more closely with the neurophysiological foundations of emotions. However, current applications of spatial-temporal graphs have not deeply explored the construction of dynamic brain region relationships under emotional states. Existing methods typically treat the graph embeddings of EEG time slices as wholes to learn temporal dependencies between slices, thus limiting spatial dependencies to within individual slices. In other words, current research mainly focuses on the temporal variations of spatial relationships within isolated segments and is unable to construct connections between different nodes across slices.

8 Future Directions

In this section, we discuss some future research directions and possible approaches for graph application in EEG-based emotion recognition in addition to the challenges or limitations mentioned in the remarks in the previous sections.

Temporal Graph: This method aims to address the issue of incomplete temporal dependency present in current spatio-temporal graph methods, as discussed in Section 7.3. This incomplete dependency refers to the lack of correlation between different channels across time slices. The spatio-temporal dependencies learned by existing methods are limited to the same electrode across different time slices, such as v_i^{t-1} and v_i^t , while ignoring the relationships between different electrodes across time slices, such as v_i^{t-1} and v_j^t . These inter-electrode relationships across time slices correspond to the delayed responses of brain regions in emotional states, indicating that interactions between brain regions exhibit asynchrony. The persistence of emotions involves asynchronous activities of brain regions, representing delayed responses of other regions to the current region’s activity [50]. Therefore, a potential direction is to establish a temporal graph between adjacent time slices, where edges exist between nodes across time slices rather than within the same time slice. In summary, computing a temporal graph for every pair of adjacent time slices can model the more intricate temporal dependencies of brain regions in emotional states.

Dependency-Level Interpretability: Currently, interpretability in graph methods for this field has garnered little attention. However, given the physiological basis of emotion-related EEG, interpretability is a crucial direction for the future. We argue that the necessary interpretability methods in this field should focus on explaining the graph structure, specifically identifying which edges, rather than just nodes, are most important for emotion labels. Most interpretability methods are feature-level, such as [9, 89] that determine which channels or time slices are most important for emotion prediction. Due to the direct relationship between emotions and brain region dependencies, a promising approach is to highlight the significant edges, as illustrated in XGNN [113], thereby reflecting crucial brain region relationships in emotion-related EEG. Another potential direction is using this method to learn interpretable sparse

graphs. Similar to XGNN, by predicting how adding edges to the current graph affects the final label and setting an upper limit on the number of edges, it is possible to identify the most impactful edges for the final emotion label. These edges represent the sparse cortical activity between brain regions during emotional activities.

Heterogeneous Graph for Mixed Emotion: Although heterogeneous graphs have seen limited application in this field, we argue they represent a future mainstream trend, particularly as a promising approach for mixed emotion recognition. It is increasingly recognized that emotions are often mixed, meaning that subjects frequently experience two or more emotions simultaneously [12, 48, 105]. The potential of heterogeneous graphs lies in their ability to reveal the blurred boundaries between mixed emotional states more effectively than brain-centric emotional interaction models alone. By integrating multimodal physiological data, heterogeneous graphs build upon EEG-based brain emotion interaction models to construct comprehensive representations of emotional interactions within brain-based physiological systems. These interactions are strongly associated with emotion; for example, there is a correlation between prefrontal cortex oxygenation and decreased facial skin blood flow during positive emotions, whereas this correlation is less pronounced during negative emotions [69]. In summary, mixed emotion recognition is a crucial future research direction in this field, and the potential of heterogeneous graphs to elucidate the boundaries between mixed emotions makes it a method deserving of greater attention.

9 Conclusion

The rapid development of graph-based methods has revolutionized EEG-based emotion recognition. In this survey, we provide a comprehensive and updated review of graph techniques specifically designed for this field. We propose a novel taxonomy based on key components such as feature selection, edge computation, node aggregation, and graph manipulation. Our survey facilitates understanding the underlying mechanisms of applying graph methods to emotion recognition. Furthermore, we believe that consolidating the latest advancements and exploring future directions will inspire more innovative works within EEG-based emotion recognition.

References

- [1] Ralph Adolphs. 2003. Cognitive neuroscience of human social behaviour. *Nature reviews neuroscience* 4, 3 (2003), 165–178.
- [2] Turky Alotaiby, Fathi E Abd El-Samie, Saleh A Alshebeili, and Ishtiaq Ahmad. 2015. A review of channel selection algorithms for EEG signal processing. *EURASIP Journal on Advances in Signal Processing* 2015 (2015), 1–21.
- [3] Shiva Asadzadeh, Tohid Yousefi Rezaii, Soosan Beheshti, and Saeed Meshgini. 2022. Accurate emotion recognition using Bayesian model based EEG sources as dynamic graph convolutional neural network nodes. *Scientific Reports* 12, 1 (2022), 10282.
- [4] Maryam Bijanzadeh, Ankit N Khambhati, Maansi Desai, Deanna L Wallace, Alia Shafii, Heather E Dawes, Virginia E Sturm, and Edward F Chang. 2022. Decoding naturalistic affective behaviour from spectro-spatial features in multiday human iEEG. *Nature human behaviour* 6, 6 (2022), 823–836.
- [5] Bob Bramson, Sjoerd Meijer, Annelies van Nuland, Ivan Toni, and Karin Roelofs. 2023. Anxious individuals shift emotion control from lateral frontal pole to dorsolateral prefrontal cortex. *Nature Communications* 14, 1 (2023), 4880.
- [6] Ed Bullmore and Olaf Sporns. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience* 10, 3 (2009), 186–198.
- [7] Hao Chao, Yiming Cao, and Yongli Liu. 2023. Multi-channel EEG emotion recognition through residual graph attention neural network. *Frontiers in Neuroscience* 17 (2023), 1135850.
- [8] Bianna Chen, CL Philip Chen, and Tong Zhang. 2024. GDDN: Graph Domain Disentanglement Network for Generalizable EEG Emotion Recognition. *IEEE Transactions on Affective Computing* (2024).
- [9] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*. PMLR, 883–892.
- [10] Cheng Cheng, Wenzhe Liu, Lin Feng, and Ziyu Jia. 2024. Dense Graph Convolutional With Joint Cross-Attention Network for Multimodal Emotion Recognition. *IEEE Transactions on Computational Social Systems* (2024).
- [11] Cheng Cheng, Zikang Yu, Yong Zhang, and Lin Feng. 2023. Hybrid Network Using Dynamic Graph Convolution and Temporal Self-Attention for EEG-Based Emotion Recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2023), 1–11. <https://doi.org/10.1109/TNNLS.2023.3319315>
- [12] Alex S Cohen, Annie St-Hilaire, Jennifer M Aakre, and Nancy M Docherty. 2009. Understanding anhedonia in schizophrenia through lexical analysis of natural speech. *Cognition and emotion* 23, 3 (2009), 569–586.
- [13] Beatrice De Gelder. 2006. Towards the neurobiology of emotional body language. *Nature reviews neuroscience* 7, 3 (2006), 242–249.
- [14] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems* 29 (2016).
- [15] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuahao Zeng, and Cuntai Guan. 2023. LGGNet: Learning from local-global-graph representations for brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems* (2023).
- [16] Guanglong Du, Jinshao Su, Linlin Zhang, Kang Su, Xueqian Wang, Shaohua Teng, and Peter Xiaoping Liu. 2022. A multi-dimensional graph convolution network for EEG emotion recognition. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–11.
- [17] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*. IEEE, 81–84.
- [18] Xiaoxu Fan, Madaline Mocchi, Bailey Pascuzzi, Jiayang Xiao, Brian A Metzger, Raissa K Mathura, Carl Hacker, Joshua A Adkinson, Eleonora Bartoli, Salma Elhassan, et al. 2024. Brain mechanisms underlying the emotion processing bias in treatment-resistant depression. *Nature Mental Health* (2024), 1–10.
- [19] Lin Feng, Cheng Cheng, Mingyan Zhao, Huiyuan Deng, and Yong Zhang. 2022. EEG-based emotion recognition using spatial-temporal graph convolutional LSTM with attention mechanism. *IEEE Journal of Biomedical and Health Informatics* 26, 11 (2022), 5406–5417.
- [20] Alex Fornito, Andrew Zalesky, and Michael Breakspear. 2015. The connectomics of brain disorders. *Nature Reviews Neuroscience* 16, 3 (2015), 159–172.
- [21] Angela D Friederici, Noam Chomsky, Robert C Berwick, Andrea Moro, and Johan J Bolhuis. 2017. Language, mind and brain. *Nature human behaviour* 1, 10 (2017), 713–722.
- [22] Dongrui Gao, Haokai Zhang, Pengrui Li, Tian Tang, Shihong Liu, Zhihong Zhou, Shaofei Ying, Ye Zhu, and Yongqing Zhang. 2024. A Local-Ascending-Global Learning Strategy for Brain-Computer Interface. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 10039–10047.
- [23] Hongxiang Gao, Xingyao Wang, Zhenghua Chen, Min Wu, Zhipeng Cai, Lulu Zhao, Jianqing Li, and Chengyu Liu. 2024. Graph Convolutional Network With Connectivity Uncertainty for EEG-Based Emotion Recognition. *IEEE Journal of Biomedical and Health Informatics* (2024).
- [24] Bruno L Giordano, Caroline Whiting, Nikolaus Kriegeskorte, Sonja A Kotz, Joachim Gross, and Pascal Belin. 2021. The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nature human behaviour* 5, 9 (2021), 1203–1213.
- [25] Peiliang Gong, Pengpai Wang, Yueying Zhou, and Daoqiang Zhang. 2023. A spiking neural network with adaptive graph convolution and lstm for eeg-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), 1440–1450.
- [26] Xue L Gong, Alexander G Huth, Fatma Deniz, Keith Johnson, Jack L Gallant, and Frédéric E Theunissen. 2023. Phonemic segmentation of narrative speech in human cerebral cortex. *Nature communications* 14, 1 (2023), 4309.
- [27] Manuel Graña and Igone Morais-Quilez. 2023. A review of Graph Neural Networks for Electroencephalography data analysis. *Neurocomputing* (2023), 126901.
- [28] Xiaoqing Gu, Weiwei Cai, Ming Gao, Yizhang Jiang, Xin Ning, and Pengjiang Qian. 2022. Multi-source domain transfer discriminative dictionary learning modeling for electroencephalogram-based emotion recognition. *IEEE Transactions on Computational Social Systems* 9, 6 (2022), 1604–1612.
- [29] Yun Gu, Xinyue Zhong, Cheng Qu, Chuannjun Liu, and Bin Chen. 2023. A domain generative graph network for EEG-based emotion recognition. *IEEE Journal of Biomedical and Health Informatics* 27, 5 (2023), 2377–2386.
- [30] Wenhui Guo and Yanjiang Wang. 2024. Convolutional gated recurrent unit-driven multidimensional dynamic graph neural network for subject-independent emotion recognition. *Expert Systems with Applications* 238 (2024), 121889.

- [31] Douglas Heaven. 2020. Why faces don't always tell the truth about feelings. *Nature* 578, 7796 (2020), 502–505.
- [32] John D Herrington, Aprajita Mohanty, Nancy S Koven, Joscelyn E Fisher, Jennifer L Stewart, Marie T Banich, Andrew G Webb, Gregory A Miller, and Wendy Heller. 2005. Emotion-modulated performance and activity in left dorsolateral prefrontal cortex. *Emotion* 5, 2 (2005), 200.
- [33] Marlies Houben, Wim Van Den Noortgate, and Peter Kuppens. 2015. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological bulletin* 141, 4 (2015), 901.
- [34] Wuliang Huang, Yiqiang Chen, Xinlong Jiang, Teng Zhang, and Qian Chen. 2023. GJFusion: A channel-level correlation construction method for multi-modal physiological signal fusion. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2023), 1–23.
- [35] Ziyu Jia, Youfang Lin, Jing Wang, Zhiyang Feng, Xiangheng Xie, and Caijie Chen. 2021. HetEmotionNet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1047–1056.
- [36] Ming Jin, Changde Du, Huiguang He, Ting Cai, and Jinpeng Li. 2024. PGCN: Pyramidal graph convolutional network for EEG emotion recognition. *IEEE Transactions on Multimedia* (2024).
- [37] Ming Jin and Jinpeng Li. 2023. Graph to Grid: Learning Deep Representations for Multimodal Emotion Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5985–5993.
- [38] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [39] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [40] Sotaro Kita and Karen Emmorey. 2023. Gesture links language and cognition for spoken and signed languages. *Nature Reviews Psychology* 2, 7 (2023), 407–420.
- [41] Dominik Klepl, Min Wu, and Fei He. 2024. Graph neural network-based eeg classification: A survey. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2024).
- [42] Hedy Kober, Lisa Feldman Barrett, Josh Joseph, Eliza Bliss-Moreau, Kristen Lindquist, and Tor D Wager. 2008. Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage* 42, 2 (2008), 998–1031.
- [43] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.
- [44] Wanpeng Kong, Min Qiu, Menghang Li, Xuanyu Jin, and Li Zhu. 2022. Causal graph convolutional neural network for emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems* 15, 4 (2022), 1686–1693.
- [45] Xiaolu Kong, Ru Kong, Csaba Orban, Peng Wang, Shaoshi Zhang, Kevin Anderson, Avram Holmes, John D Murray, Gustavo Deco, Martijn van den Heuvel, et al. 2021. Sensory-motor cortices shape functional connectivity dynamics in the human brain. *Nature communications* 12, 1 (2021), 6373.
- [46] Philip A Kragel and Kevin S LaBar. 2016. Decoding the nature of emotion in the brain. *Trends in cognitive sciences* 20, 6 (2016), 444–455.
- [47] Pavitra Krishnaswamy, Gabriel Obregon-Henao, Jyrki Ahveninen, Sheraz Khan, Behash Babadi, Juan Eugenio Iglesias, Matti S Hämäläinen, and Patrick L Purdon. 2017. Sparsity enables estimation of both subcortical and cortical activity from MEG and EEG. *Proceedings of the National Academy of Sciences* 114, 48 (2017), E10465–E10474.
- [48] Jeff T Larsen and A Peter McGraw. 2011. Further evidence for mixed emotions. *Journal of personality and social psychology* 100, 6 (2011), 1095.
- [49] Giada Lettieri, Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Paolo Papale, Monica Bettà, Pietro Pietrini, and Luca Cecchetti. 2019. Emotionotopy in the human right temporo-parietal cortex. *Nature communications* 10, 1 (2019), 5568.
- [50] Marc D Lewis. 2005. Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and brain sciences* 28, 2 (2005), 169–194.
- [51] Cunbo Li, Peiyang Li, Yangsong Zhang, Ning Li, Yajing Si, Fali Li, Zehong Cao, Huafu Chen, Badong Chen, Dezhong Yao, et al. 2023. Effective emotion recognition by learning discriminative graph topologies in EEG brain networks. *IEEE Transactions on Neural Networks and Learning Systems* (2023).
- [52] Cunbo Li, Tian Tang, Yue Pan, Lei Yang, Shuhan Zhang, Zhaojin Chen, Peiyang Li, Dongrui Gao, Huafu Chen, Fali Li, et al. 2024. An Efficient Graph Learning System for Emotion Recognition Inspired by the Cognitive Prior Graph of EEG Brain Network. *IEEE Transactions on Neural Networks and Learning Systems* (2024).
- [53] Chao Li, Feng Wang, Ziping Zhao, Haishuai Wang, and Björn W Schuller. 2024. Attention-based Temporal Graph Representation Learning for EEG-based Emotion Recognition. *IEEE Journal of Biomedical and Health Informatics* (2024).
- [54] Jingcong Li, Shuqi Li, Jiahui Pan, and Fei Wang. 2021. Cross-subject EEG emotion recognition with self-organized graph neural network. *Frontiers in Neuroscience* 15 (2021), 611653.
- [55] Qilin Li, Tong Zhang, CL Philip Chen, Ke Yi, and Long Chen. 2022. Residual GCB-Net: Residual graph convolutional broad network on emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems* 15, 4 (2022), 1673–1685.
- [56] Rui Li, Chao Ren, Yiqing Ge, Qiqi Zhao, Yikun Yang, Yuhan Shi, Xiaowei Zhang, and Bin Hu. 2023. MTLFuseNet: a novel emotion recognition model based on deep latent feature fusion of EEG signals and multi-task learning. *Knowledge-Based Systems* 276 (2023), 110756.
- [57] Rui Li, Yiting Wang, and Bao-Liang Lu. 2021. A multi-domain adaptive graph convolutional network for EEG-based emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5565–5573.
- [58] Wei Li, Hong Wang, and Luhe Zhuang. 2023. GCNs-FSMI: EEG recognition of mental illness based on fine-grained signal features and graph mutual information maximization. *Expert Systems With Applications* 228 (2023), 120227.
- [59] Wei Li, Mingming Wang, Junyi Zhu, and Aiguo Song. 2023. EEG-based emotion recognition using trainable adjacency relation driven graph convolutional network. *IEEE Transactions on Cognitive and Developmental Systems* 15, 4 (2023), 1656–1672.
- [60] Xiaojun Li, CL Philip Chen, Bianna Chen, and Tong Zhang. 2024. Gusa: Graph-based unsupervised subdomain adaptation for cross-subject eeg emotion recognition. *IEEE Transactions on Affective Computing* (2024).
- [61] Xiaoxu Li, Wenming Zheng, Yuan Zong, Hongli Chang, and Cheng Lu. 2021. Attention-based spatio-temporal graphic lstm for eeg emotion recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
- [62] Yang Li, Ji Chen, Fu Li, Boxun Fu, Hao Wu, Youshuo Ji, Yijin Zhou, Yi Niu, Guangming Shi, and Wenming Zheng. 2022. GMSS: Graph-based multi-task self-supervised learning for EEG emotion recognition. *IEEE Transactions on Affective Computing* 14, 3 (2022), 2512–2525.
- [63] Yan Li, Liang Zhang, Xiangyuan Lan, and Dongmei Jiang. 2023. Towards adaptable graph representation learning: An adaptive multi-graph contrastive transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6063–6071.
- [64] Zhongjie Li, Gaoyan Zhang, Longbiao Wang, Jianguo Wei, and Jianwu Dang. 2023. Emotion recognition using spatial-temporal EEG features through convolutional graph attention network. *Journal of Neural Engineering* 20, 1 (2023), 016046.
- [65] Kristen A Lindquist, Tor D Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. 2012. The brain basis of emotion: a meta-analytic review. *Behavioral and brain sciences* 35, 3 (2012), 121–143.
- [66] Chenyu Liu, xinliang Zhou, Jiaping Xiao, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. 2024. VSGT: Variational Spatial and Gaussian Temporal Graph Models for EEG-based Emotion Recognition. In *IJCAI*.
- [67] Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. [n. d.]. VBH-GNN: Variational Bayesian Heterogeneous Graph Neural Networks for Cross-subject Emotion Recognition. In *The Twelfth International Conference on Learning Representations*.
- [68] Hanjie Liu, Jinren Zhang, Qingshan Liu, and Jinde Cao. 2022. Minimum spanning tree based graph neural network for emotion classification using EEG. *Neural Networks* 145 (2022), 308–318.
- [69] Kanji Matsukawa, Ryota Asahara, Miho Yoshikawa, and Kana Endo. 2018. Deactivation of the prefrontal cortex during exposure to pleasantly-charged emotional challenge. *Scientific Reports* 8, 1 (2018), 14540.
- [70] Mason McClay, Matthew E Sachs, and David Clewett. 2023. Dynamic emotional states shape the episodic structure of memory. *Nature Communications* 14, 1 (2023), 6533.
- [71] Deng Pan, Haohao Zheng, Feifan Xu, Yu Ouyang, Zhe Jia, Chu Wang, and Hong Zeng. 2023. MSFR-GCN: A multi-scale feature reconstruction graph convolutional network for EEG emotion and cognition recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2023).
- [72] Jiahui Pan, Rongming Liang, Zhipeng He, Jingcong Li, Yan Liang, Xinjie Zhou, Yanbin He, and Yuanqiang Li. 2023. ST-SCGNN: a spatio-temporal self-constructing graph neural network for cross-subject EEG-based emotion recognition and consciousness detection. *IEEE Journal of Biomedical and Health Informatics* (2023).
- [73] Tongjie Pan, Yalan Ye, Hecheng Cai, Shudong Huang, Yang Yang, and Guoqing Wang. 2023. Multimodal physiological signals fusion for online emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5879–5888.
- [74] Tongjie Pan, Yalan Ye, Yangwuyong Zhang, Kunshu Xiao, and Hecheng Cai. 2024. Online multi-hypergraph fusion learning for cross-subject emotion recognition. *Information Fusion* 108 (2024), 102338.
- [75] Anand S Pandit, Paul Expert, Renaud Lambiotte, Valerie Bonnelle, Robert Leech, Federico E Turkheimer, and David J Sharp. 2013. Traumatic brain injury impairs small-world topology. *Neurology* 80, 20 (2013), 1826–1833.
- [76] James C Pang, Kevin M Aquino, Marianne Oldehinkel, Peter A Robinson, Ben D Fulcher, Michael Breakspear, and Alex Fornito. 2023. Geometric constraints on human brain function. *Nature* (2023), 1–9.
- [77] Dan Peng, Wei-Long Zheng, Luyu Liu, Wei-Bang Jiang, Ziyi Li, Yong Lu, and Bao-Liang Lu. 2023. Identifying sex differences in EEG-based emotion recognition using graph convolutional network with attention mechanism. *Journal of Neural Engineering* 20, 6 (2023), 066010.

- [78] Yong Peng, Fengzhe Jin, Wanpeng Kong, Feiping Nie, Bao-Liang Lu, and Andrzej Cichocki. 2022. OGSSL: A semi-supervised classification model coupled with optimal graph learning for EEG emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), 1288–1297.
- [79] Yong Peng, Wanpeng Kong, Feiwei Qin, Feiping Nie, Jinglong Fang, Bao-Liang Lu, and Andrzej Cichocki. 2021. Self-weighted semi-supervised classification for joint EEG-based emotion recognition and affective activation patterns mining. *IEEE Transactions on Instrumentation and Measurement* 70 (2021), 1–11.
- [80] Yong Peng, Feiwei Qin, Wanpeng Kong, Yuan Ge, Feiping Nie, and Andrzej Cichocki. 2021. GFIL: A unified framework for the importance analysis of features, frequency bands, and channels in EEG-based emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems* 14, 3 (2021), 935–947.
- [81] Yong Peng, Wenjuan Wang, Wanpeng Kong, Feiping Nie, Bao-Liang Lu, and Andrzej Cichocki. 2022. Joint feature adaptation and graph adaptive label propagation for cross-subject emotion recognition from EEG signals. *IEEE Transactions on Affective Computing* 13, 4 (2022), 1941–1958.
- [82] Luiz Pessoa. 2008. On the relationship between emotion and cognition. *Nature reviews neuroscience* 9, 2 (2008), 148–158.
- [83] Luiz Pessoa and Ralph Adolphs. 2010. Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nature reviews neuroscience* 11, 11 (2010), 773–782.
- [84] Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. 2011. Functional network organization of the human brain. *Neuron* 72, 4 (2011), 665–678.
- [85] Xiangkai Qiu, Shenglin Wang, Ruqing Wang, Yiling Zhang, and Liya Huang. 2023. A multi-head residual connection GCN for EEG emotion recognition. *Computers in Biology and Medicine* 163 (2023), 107126.
- [86] Rebecca D Ray and David H Zald. 2012. Anatomical insights into the interaction of emotion and cognition in the prefrontal cortex. *Neuroscience & Biobehavioral Reviews* 36, 1 (2012), 479–501.
- [87] Raymond Salvador, John Suckling, Martin R Coleman, John D Pickard, David Menon, and ED Bullmore. 2005. Neurophysiological architecture of functional magnetic resonance images of human brain. *Cerebral cortex* 15, 9 (2005), 1332–1342.
- [88] C Daniel Salzman and Stefano Fusi. 2010. Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. *Annual review of neuroscience* 33, 1 (2010), 173–202.
- [89] Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. 2020. Interpreting graph neural networks for NLP with differentiable edge masking. *arXiv preprint arXiv:2010.00577* (2020).
- [90] Harald T Schupp, Markus Junghöfer, Almut I Weike, and Alfons O Hamm. 2003. Attention and emotion: an ERP analysis of facilitated emotional stimulus processing. *Neuroreport* 14, 8 (2003), 1107–1110.
- [91] Martin I Sereno and Ruey-Song Huang. 2006. A human parietal face area contains aligned head-centered visual and tactile maps. *Nature neuroscience* 9, 10 (2006), 1337–1343.
- [92] Miseon Shim, Chang-Hwan Im, Yong-Wook Kim, and Seung-Hwan Lee. 2018. Altered cortical functional network in major depressive disorder: A resting-state electroencephalogram study. *NeuroImage: Clinical* 19 (2018), 1000–1007.
- [93] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* 3, 1 (2011), 42–55.
- [94] Tengfei Song, Suyuan Liu, Wenming Zheng, Yuan Zong, and Zhen Cui. 2020. Instance-adaptive graph for EEG emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 2701–2708.
- [95] Tengfei Song, Suyuan Liu, Wenming Zheng, Yuan Zong, Zhen Cui, Yang Li, and Xiaoyan Zhou. 2021. Variational instance-adaptive graph for EEG emotion recognition. *IEEE Transactions on Affective Computing* 14, 1 (2021), 343–356.
- [96] Tengfei Song, Wenming Zheng, Suyuan Liu, Yuan Zong, Zhen Cui, and Yang Li. 2021. Graph-embedded convolutional neural network for image-based EEG emotion recognition. *IEEE Transactions on Emerging Topics in Computing* 10, 3 (2021), 1399–1413.
- [97] Tengfei Song, Wenming Zheng, Cheng Lu, Yuan Zong, Xilei Zhang, and Zhen Cui. 2019. MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access* 7 (2019), 12177–12191.
- [98] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing* 11, 3 (2018), 532–541.
- [99] Mingyi Sun, Weigang Cui, Shuyue Yu, Hongbin Han, Bin Hu, and Yang Li. 2022. A dual-branch dynamic graph convolution based adaptive transformer feature fusion network for EEG emotion recognition. *IEEE Transactions on Affective Computing* 13, 4 (2022), 2218–2228.
- [100] Arielle Tambini, Ulrike Rimmele, Elizabeth A Phelps, and Lila Davachi. 2017. Emotional brain states carry over and enhance future memory formation. *Nature neuroscience* 20, 2 (2017), 271–278.
- [101] Hao Tang, Songyun Xie, Xinzhou Xie, Yujie Cui, Bohan Li, Dalu Zheng, Yu Hao, Xiangming Wang, Yiye Jiang, and Zhongyu Tian. 2024. Multi-Domain Based Dynamic Graph Representation Learning for EEG Emotion Recognition. *IEEE Journal of Biomedical and Health Informatics* (2024).
- [102] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* 3, 1 (2021), 42–50.
- [103] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [104] Yiming Wang, Bin Zhang, and Lamei Di. 2024. Research Progress of EEG-Based Emotion Recognition: A Survey. *Comput. Surveys* 56, 11 (2024), 1–49.
- [105] Patti Williams and Jennifer L Aaker. 2002. Can mixed emotions peacefully coexist? *Journal of consumer research* 28, 4 (2002), 636–649.
- [106] Mengqi Wu, CL Philip Chen, Bianna Chen, and Tong Zhang. 2024. Grop: Graph Orthogonal Purification Network for EEG Emotion Recognition. *IEEE Transactions on Affective Computing* (2024).
- [107] Tao Xu, Wang Dang, Jiaobao Wang, and Yun Zhou. 2023. DAGAM: a domain adversarial graph attention model for subject-independent EEG-based emotion recognition. *Journal of Neural Engineering* 20, 1 (2023), 016022.
- [108] Yunlong Xue, Wenming Zheng, Yuan Zong, Hongli Chang, and Xingxun Jiang. 2022. Adaptive hierarchical graph convolutional network for eeg emotion recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
- [109] Huachao Yan, Kailing Guo, Xiaofen Xing, and Xiangmin Xu. 2024. Bridge Graph Attention based Graph Convolution Network with Multi-Scale Transformer for EEG Emotion Recognition. *IEEE Transactions on Affective Computing* (2024).
- [110] Kun Yang, Zhenning Yao, Keze Zhang, Jing Xu, Li Zhu, Shichao Cheng, and Jianhai Zhang. 2024. Automatically Extracting and Utilizing EEG Channel Importance Based on Graph Convolutional Network for Emotion Recognition. *IEEE Journal of Biomedical and Health Informatics* (2024).
- [111] Mengqing Ye, CL Philip Chen, and Tong Zhang. 2022. Hierarchical dynamic graph convolutional network with interpretability for EEG-based emotion recognition. *IEEE transactions on neural networks and learning systems* (2022).
- [112] Weishan Ye, Zhiqiu Zhang, Fei Teng, Min Zhang, Jianhong Wang, Dong Ni, Fali Li, Peng Xu, and Zhen Liang. 2024. Semi-supervised dual-stream self-attentive adversarial graph contrastive learning for cross-subject eeg-based emotion recognition. *IEEE Transactions on Affective Computing* (2024).
- [113] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. 2020. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 430–438.
- [114] Hong Zeng, Qi Wu, Yanping Jin, Haohao Zheng, Mingming Li, Yue Zhao, Hua Hu, and Wanpeng Kong. 2022. Siam-GCAN: A Siamese graph convolutional attention network for EEG emotion recognition. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–9.
- [115] Guanhua Zhang, Minjing Yu, Yong-Jin Liu, Guozhen Zhao, Dan Zhang, and Wenming Zheng. 2021. SparseDGCNN: Recognizing emotion from multichannel EEG signals. *IEEE Transactions on Affective Computing* 14, 1 (2021), 537–548.
- [116] Tong Zhang, Xuehan Wang, Xiangmin Xu, and CL Philip Chen. 2019. GCB-Net: Graph convolutional broad network and its application in emotion recognition. *IEEE Transactions on Affective Computing* 13, 1 (2019), 379–388.
- [117] Zhongyi Zhang, Qinghai Meng, LiCheng Jin, Hanguang Wang, and Huirang Hou. 2024. A novel EEG-based graph convolution network for depression detection: incorporating secondary subject partitioning and attention mechanism. *Expert Systems with Applications* 239 (2024), 122356.
- [118] Huijuan Zhao, Jingjin Liu, Zhenqian Shen, and Jingwen Yan. 2022. SCC-MPGCN: self-attention coherence clustering based on multi-pooling graph convolutional network for EEG emotion recognition. *Journal of Neural Engineering* 19, 2 (2022), 026051.
- [119] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development* 7, 3 (2015), 162–175.
- [120] Peixiang Zhong, Di Wang, and Chunyan Miao. 2020. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1290–1301.
- [121] Yijin Zhou, Fu Li, Yang Li, Youshuo Ji, Guangming Shi, Wenming Zheng, Lijian Zhang, Yuanfang Chen, and Rui Cheng. 2023. Progressive graph convolution network for EEG emotion recognition. *Neurocomputing* 544 (2023), 126262.
- [122] Vadim Zotev, Ahmad Mayeli, Masaya Misaki, and Jerzy Bodurka. 2020. Emotion self-regulation training in major depressive disorder using simultaneous real-time fMRI and EEG neurofeedback. *NeuroImage: Clinical* 27 (2020), 102331.

A Appendix

In this section, we provide a detailed categorized table as shown in Table 1.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Table 1: Summary of graph-related methods for EEG-based emotion recognition

Baseline	Feature Selection		Edge Computation				Node Aggregation		Graph Manipulation			Year	
	Temporal Nodes	Frequency Nodes	Model-independent Edge		Model-dependent Edge		Spectral-based	Spatial-based	Hierarchical Graph	Multi-Graph	Spatial-Temporal Graph		
			Prior	Signal Correlation	Weighted	Subspace							
DGCNN [98]	Raw	DE,PSD,DASM,RASM,DCAU		Gaussian Kernel			Polynomial					2018	
GCB-Net [116]		DE,PSD,DASM,RASM		Gaussian Kernel			Polynomial					2019	
IAG [94]		DE							Brain Region Graph			2020	
RGN [120]		DE					Polynomial					2020	
V-IAG [95]		DE,PSD,DASM,RASM					Polynomial					2021	
SparseDGNN [115]	ASM	DE,PSD,DASM,RASM,DCAU		Gaussian Kernel			Polynomial					2021	
HefEmotionNet [35]	Raw	DE,PSD		Mutual Information			Linear					2021	
ASTG-LSTM [61]		DE,PSD					Polynomial					2021	
MD-AGCN [57]		DE					Linear					2021	
GECNN [96]		DE,PSD,IHS					Linear					2021	
SOGN [54]	ASM	DE,PSD,DASM,RASM,DCAU					Linear		Brain Region Graph			2021	
GFIL [80]		DE,PSD,DASM,RASM,DCAU					Linear					2021	
SWSC [79]		DE,PSD,DASM,RASM,DCAU					Linear					2021	
Siam-GCAN [114]		DE					Linear					2022	
Residual GCB-Net [55]	Raw	DE,PSD,DASM,RASM					Polynomial					2022	
SCC-MPGCN [118]		DE					Polynomial					2022	
OGSSL [78]		DE					Polynomial					2022	
MGSTGNN [68]		DE					Polynomial					2022	
JAGP [81]		DE					Polynomial					2022	
HD-GCN [111]		DE					Polynomial					2022	
GMSS [62]		DE					Polynomial					2022	
TARDGCN [59]		DE,Welch					Polynomial					2022	
AHG-GCN [108]		DE					Polynomial					2022	
BLB-DGCNN [3]	Raw	DE					Linear					2022	
MD-GCN [16]		DE					Linear					2022	
DIGC-ATPNNet-ATFL [99]		DE,PSD					Linear					2022	
CGCNN [44]		DE,PSD,DASM					Linear					2022	
MDTDDL [28]		DE,PSD					Linear					2022	
OMIGL [73]		DE,PSD					Linear					2023	
LGGNet [15]	Raw	DE,PSD					Linear					2023	
G2G [37]		DE,PSD					Linear					2023	
STFCGAT [64]		DE					Attention					2023	
MESNP [51]		PSD					Attention					2023	
ST-GCLSTM [19]		DE					Attention					2023	
DAGAM [107]		DE					Attention					2023	
DGNN [29]		DE					Attention					2023	
MSFR-GCN [71]		DE					Attention					2023	
ST-SCGNN [72]		DE,STFT					Attention					2023	
SGLNet [25]	Raw	DE					Attention					2023	
GIFusion [34]		DE					Attention					2023	
AMGCT [63]		DE					Attention					2023	
ASGC [77]		DE					Attention					2023	
ST-SCGNN [72]		DE					Attention					2023	
PGCN [121]		DE,STFT					Attention					2023	
HN-DGST [11]		DE,PSD,DASM,RASM,DCAU					Attention					2023	
MRGCN [85]		DE					Attention					2023	
ResGAT [7]	Raw	DE					Attention					2023	
MTFuseNet [56]		DE					Attention					2023	
GCNs-FSMI [58]	Raw	DE					Attention					2023	
LAG [22]		DE					Attention					2024	
BF-GCN [52]		DE					Attention					2024	
PGCN [36]		DE,STFT					Attention					2024	
Gusa [60]		DE,PSD,DASM,RASM,DCAU					Attention					2024	
GDDN [8]		DE,STFT					Attention					2024	
BGAGCN [109]	Raw	DE,PSD					Attention					2024	
DG-JCA [10]		DE,PSD,DASM,RASM,DCAU					Attention					2024	
CU-GCN [23]		DE,PSD,DASM,RASM					Attention					2024	
ATGRNet [53]		DE,PSD					Attention					2024	
MP^2GRL [101]		DE,PSD					Attention					2024	
CCSR-GCN [110]		DE					Attention					2024	
Grop [106]		DE					Attention					2024	
DS-AGC [112]		DE					Attention					2024	
OmMHP [74]	Raw	PSD					Attention					2024	
GRU-MIDGN [30]	Raw	DE					Attention					2024	
VSGT [66]	Raw	Raw					Attention					2024	
SSPA-GCN [117]	Raw	Raw					Attention					2024	
VBH-GNN [67]	Raw	Raw					Attention					2024	

EEG-SCMM: Soft Contrastive Masked Modeling for Cross-Corpus EEG-Based Emotion Recognition

Qile Liu^{1,2}, Weishan Ye^{1,2}, Yulu Liu^{1,2}, Zhen Liang^{1,2,*}

¹School of Biomedical Engineering, Shenzhen University, Shenzhen, 518060, Guangdong, China

²Guangdong Provincial Key Laboratory of Biomedical Measurements and Ultrasound Imaging, Shenzhen, China
 {liuqile2022, 2110246024, 2021220004}@email.szu.edu.cn, janezliang@szu.edu.cn

Abstract

Emotion recognition using electroencephalography (EEG) signals has garnered widespread attention in recent years. However, existing studies have struggled to develop a sufficiently generalized model suitable for different datasets without re-training (**cross-corpus**). This difficulty arises because distribution differences across datasets far exceed the intra-dataset variability. To solve this problem, we propose a novel Soft Contrastive Masked Modeling (**SCMM**) framework. Inspired by emotional continuity, SCMM integrates soft contrastive learning with a new hybrid masking strategy to effectively mine the "short-term continuity" characteristics inherent in human emotions. During the self-supervised learning process, soft weights are assigned to sample pairs, enabling adaptive learning of similarity relationships across samples. Furthermore, we introduce an aggregator that weightedly aggregates complementary information from multiple close samples based on pairwise similarities among samples to enhance fine-grained feature representation, which is then used for original sample reconstruction. Extensive experiments on the SEED, SEED-IV and DEAP datasets show that SCMM achieves state-of-the-art (SOTA) performance, outperforming the second-best method by an average accuracy of 4.26% under two types of cross-corpus conditions (same-class and different-class) for EEG-based emotion recognition.

Introduction

Emotions are human attitudinal experiences and behavioral responses to objective things, closely related to health conditions and behavioral patterns (Wang, Zhang, and Tang 2024). Compared to speech (Singh and Goel 2022), gestures (Noroozi et al. 2018), and facial expressions (Canal et al. 2022), electroencephalography (EEG) offers a more direct and objective measurement of human emotions by capturing brain activity across various scalp locations (Hu et al. 2019). Recently, researchers have increasingly emphasised EEG-based emotion recognition (Zhong, Wang, and Miao 2020; Zhao, Yan, and Lu 2021; Zhang, Liu, and Zhong 2022; Gao et al. 2024), aiming to advance the development of affective brain-computer interfaces (ABCIs). However, three critical challenges remain to be addressed in current approaches.

(1) **Dataset Specificity.** Most existing EEG-based emotion recognition methods are typically designed for a single dataset, necessitating model retraining when the dataset changes. This requirement significantly limits the model's generalizability and scalability, hindering its application across different datasets. To tackle this issue, the concept of **cross-corpus** has been proposed, which is designed to be generalized across multiple datasets. A cross-corpus model is trained on one dataset and can be directly applied to another without the need for retraining from scratch. This concept, which originated in natural language processing (Schuller et al. 2010; Zhang et al. 2011), has been extended to various domains in recent years (Rayatdoost and Soleymani 2018; Chien, Yang, and Lee 2020; Ryumina, Dresvyanskiy, and Karpov 2022). Although existing EEG-based emotion recognition methods, such as BiDANN (Li et al. 2018), TANN (Li et al. 2021), and PR-PL (Zhou et al. 2023), have achieved superior performance in within-subject or cross-subject tasks, their efficacy significantly degrades in cross-corpus scenarios, where differences in data distribution across datasets far exceed the variability within a single dataset (Rayatdoost and Soleymani 2018).

(2) **Data Availability.** Current approaches for cross-domain or cross-corpus EEG-based emotion recognition rely heavily on domain adaptation techniques, which depend extensively on the availability of labeled source data and unlabeled target data. For example, AD-TCN (He, Zhong, and Pan 2022) learned an asymmetric mapping that adapts the target domain feature encoder to the source domain, aiming to eliminate the complicated steps of target domain labeling and improve the model performance in cross-domain scenarios. Similarly, E²STN (Zhou et al. 2024) integrated content information from the source domain with style information from the target domain to create stylized emotional EEG representations. However, these methods require prior access to all labeled source data and unlabeled target data for model training, presenting a significant limitation due to data availability constraints (Liu et al. 2024a).

(3) **Ignorance of Emotional Continuity.** Unlike domain adaptation techniques, contrastive learning (CL) achieves superior performance without relying on labeled data, and has demonstrated significant potential in enhancing feature representation capabilities across various domains (Chen et al. 2020; Radford et al. 2021; Zhang et al. 2022). Current

*Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

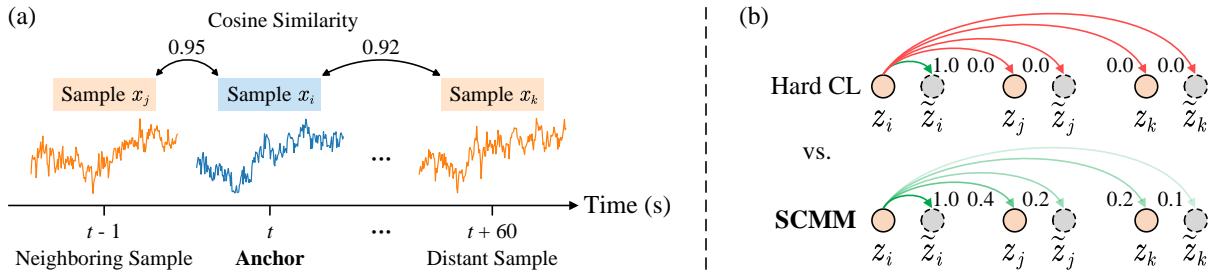


Figure 1: (a) **An illustration of emotional continuity.** We take the sample x_i at second t within an EEG trial as the anchor, and calculate the cosine similarity between x_i and its neighboring sample x_j , as well as the distant sample x_k . High cosine similarities indicate that human emotions remain relatively stable and similar over a certain period. (b) **Hard CL vs. SCMM.** Traditional hard contrastive learning considers the embeddings z_i and \tilde{z}_i of the same sample x_i and its augmented view \tilde{x}_i as positive pairs, while different samples and their augmented views are treated as negative pairs. When computing the contrastive loss, the weights for positives and negatives are set to 1 and 0, respectively. In contrast, the soft contrastive learning strategy designed in SCMM generates soft assignments for different sample pairs based on their distances in the original data space.

CL-based methods for EEG-based emotion recognition consider an anchor and its augmented views as positive pairs, while treating all other samples as negatives, as shown in Fig. 1(b) (Hard CL). For example, CLISA (Shen et al. 2022) employed contrastive learning to minimize the inter-subject differences by maximizing the similarity in EEG representations across subjects. JCFA (Liu et al. 2024a) performed joint contrastive learning across three domains to align the time- and frequency-based embeddings of the same sample in the latent time-frequency space, achieving state-of-the-art (SOTA) performance in cross-corpus scenarios. However, psychological and neuroscientific studies have shown that emotions exhibit significant "short-term continuity" characteristics (Davidson 1998; Houben, Van Den Noortgate, and Kuppens 2015). In other words, human emotions are relatively stable over certain periods, with sudden changes being rare. As illustrated in Fig. 1(a), a high cosine similarity is maintained between an anchor sample x_i and its neighboring sample x_j , and even a distant sample x_k separated by extended periods (e.g., 60 seconds). Given this nature of emotions, we propose that the definition of positive pairs in CL-based EEG analysis should extend beyond just the anchor and its augmented views. Instead, it should include a broader range of similar samples, especially those that are temporally proximal, as shown in Fig. 1(b) (SCMM). However, existing methods like JCFA (Liu et al. 2024a), which follow the traditional CL paradigm (Chen et al. 2020), may incorrectly pull apart similar but not identical samples, thus failing to capture the inherent correlations of EEG signals.

To address the aforementioned three critical issues, we propose a novel **Soft Contrastive Masked Modeling (SCMM)** framework for cross-corpus EEG-based emotion recognition. Unlike traditional hard CL shown in Fig. 1(b), SCMM considers emotional continuity and incorporates soft assignments of sample pairs. This approach enables the model to identify the fine-grained relationships between samples in a self-supervised manner, thereby enhancing the generalizability of EEG representations. Comprehensive experiments on three well-recognized datasets demonstrate

that SCMM consistently achieves SOTA performance, highlighting its superior capability and stability. In summary, the main contributions of SCMM are outlined as follows:

- Inspired by the nature of emotions, we propose a novel SCMM framework to address cross-corpus EEG-based emotion recognition. This approach assigns soft weights to sample pairs during contrastive learning to capture the similarity relationships between different samples. As a result, better feature representations of EEG signals are learned in a self-supervised manner. To the best of our knowledge, this is the first study to introduce soft contrastive learning into EEG-based emotion recognition.
- We also develop a new hybrid masking strategy to generate diverse masked samples by considering both channel and feature relationships, which is essential for enhancing contrastive learning. Additionally, we introduce an aggregator that weightedly aggregates complementary information from the embeddings of multiple close samples, enabling fine-grained feature learning and improving the model's overall capability.
- We conduct extensive experiments on three well-known datasets (SEED, SEED-IV, and DEAP), showing that SCMM achieves consistent SOTA performance compared to eight competitive methods, surpassing the second-best method by an average accuracy of 4.26%.

Methodology

Problem Definition

Given an unlabeled pre-training EEG emotion dataset $\mathcal{X} = \{x_i\}_{i=1}^N$ with N samples, where each sample $x_i \in \mathbb{R}^{C \times F}$ contains C channels and F -dimensional features, the goal is to learn a nonlinear embedding function f_θ . This function is designed to map x_i to its representation h_i that best describes itself by leveraging the emotional continuity inherent in EEG signals. Ultimately, the pre-trained model is capable of producing generalizable EEG representations that can be effectively used across different EEG emotion datasets.

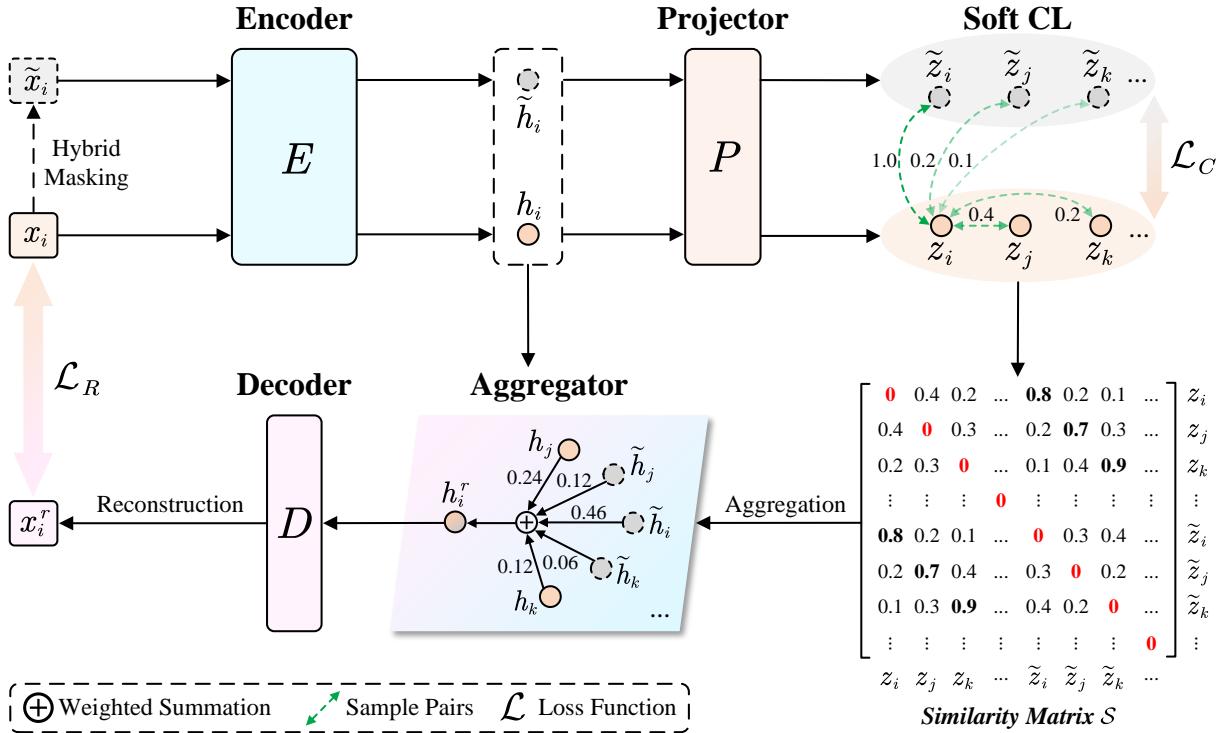


Figure 2: **The overall framework of SCMM.** The pre-training process of SCMM involves three modules: (1) hybrid masking, (2) soft contrastive learning, and (3) aggregate reconstruction.

Model Architecture

The overall framework of SCMM is illustrated in Fig. 2, which includes three main modules: hybrid masking, soft contrastive learning, and aggregate reconstruction. Below, we detail the specific design of each module and the pre-training process of SCMM.

Hybrid Masking The selection of masking strategies is crucial for CL and masked modeling (Zhang et al. 2024; Liu et al. 2024b). For an input EEG sample $x_i \in \mathcal{X}$, most existing methods use random masking (Zhang, Liu, and Zhong 2022) or channel masking (Li et al. 2022) to generate the masked sample \tilde{x}_i . The random masking strategy masks samples along the feature dimension, ignoring the inter-channel relationships of EEG signals. While a large masking ratio (e.g., 75%) can mask entire portions of certain channels, it complicates the modeling process due to significant information loss. Conversely, the channel masking strategy masks features across all dimensions of the selected channels, losing the relationships between different dimensional features. Neither approach captures both channel and feature relationships simultaneously. Therefore, we develop a new hybrid masking strategy to generate diverse masked samples by considering both channel and feature relationships.

Specifically, we first generate a random masking matrix $\text{Mask}_R \in \{0, 1\}$ with dimensions $C \times F$, and a channel masking matrix $\text{Mask}_C \in \{0, 1\}$ with dimensions $C \times F$, both derived from binomial distributions with masking ratios $r \in (0, 1)$. Here, the element values in each row of Mask_C

are either all 1s or all 0s. Next, we generate a probability matrix $U \in [0, 1]$ with dimensions $C \times 1$ for hybrid masking, which is drawn from a uniform distribution. The hybrid masking process is defined as:

$$\tilde{x}_{i,c} = \begin{cases} x_{i,c} \odot \text{Mask}_{R,c} & \text{if } \mu < U_c \leq 1 \\ x_{i,c} \odot \text{Mask}_{C,c} & \text{if } 0 \leq U_c \leq \mu \end{cases}, \quad (1)$$

where $x_{i,c}$ represents the c -th channel of x_i , and $\tilde{x}_{i,c}$ is the corresponding masked sample. \odot denotes element-wise multiplication. U_c is the probability value in the c -th row, and μ is a probability threshold that controls the weights of the two masking strategies. By integrating the hybrid masking strategy in SCMM, we enhance the diversity of masked samples, encouraging the model to learn richer and more robust feature representations that account for both channel and feature relationships within EEG signals. Figure 3 illustrates the differences between three masking strategies.

Soft Contrastive Learning Traditional hard CL treats augmented views generated from the same sample as positive pairs, and those from different samples as negative pairs (Chen et al. 2020; Eldele et al. 2021; Yue et al. 2022). During the computation of the contrastive loss, hard values (1 or 0) are assigned to sample pairs, as illustrated in Fig. 1(b) (Hard CL). However, we argue that this approach fails to account for the “short-term continuity” characteristic inherent in human emotions, leading to inaccurate modeling of inter-sample relationships and hindering the generalizability of the learned embeddings.

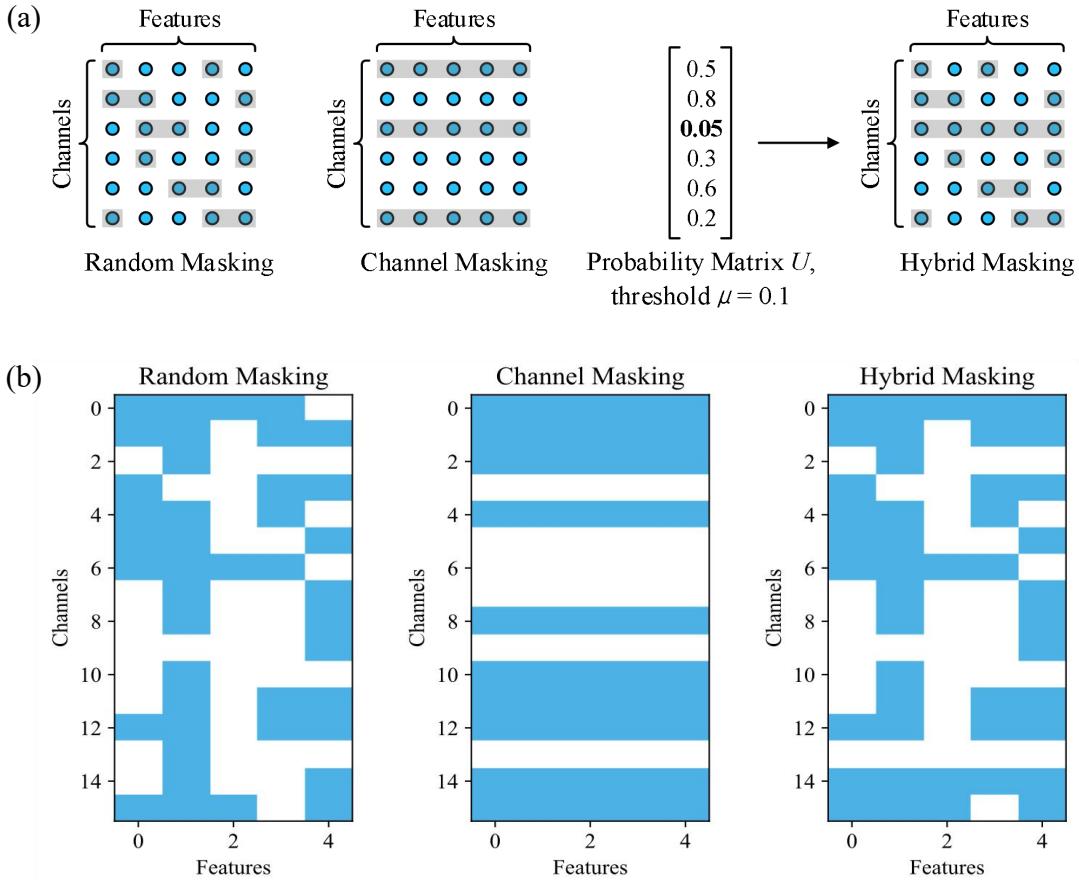


Figure 3: **Illustrations of different masking strategies.** (a) compares various masking strategies, and (b) presents examples of generated masked samples using three strategies. The masking ratio and threshold are set to $r = 0.5$ and $\mu = 0.1$, respectively.

To address this issue, we propose defining soft assignments for different sample pairs, as shown in Fig. 1(b) (SCMM). We first input x_i and \tilde{x}_i into an encoder E that maps samples to embeddings, denoted as $h_i = E(x_i)$ and $\tilde{h}_i = E(\tilde{x}_i)$. These embeddings are then projected into a latent space \mathcal{Z} using a projector P , resulting in $z_i = P(h_i)$ and $\tilde{z}_i = P(\tilde{h}_i)$. Next, we perform soft contrastive learning in \mathcal{Z} using z_i and \tilde{z}_i . Specifically, for a given pair of samples (x_i, x_j) , we first calculate the normalized distance $D(x_i, x_j)$ between x_i and x_j in the original data space as:

$$D(x_i, x_j) = \text{Norm}(\text{Dist}(x_i, x_j)) \in [0, 1], \quad (2)$$

where $\text{Dist}(\cdot, \cdot)$ is a metric function used to measure the distance between sample pairs, and $\text{Norm}(\cdot)$ denotes min-max normalization. In the experiments, we take the negative of cosine similarity as the metric function. Based on the normalized distance $D(x_i, x_j)$, we then define a soft assignment $w(x_i, x_j)$ for each pair of samples (x_i, x_j) using the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$:

$$w(x_i, x_j) = 2\alpha \cdot \sigma(-D(x_i, x_j)/\tau_w), \quad (3)$$

where $\alpha \in [0, 1]$ is a boundary parameter that controls the upper bound of soft assignments. τ_w is a sharpness parameter, where smaller values of τ_w result in greater differences

in $w(\cdot, \cdot)$ between sample pairs, and vice versa. Figure 4 depicts heat maps of soft assignments $w(\cdot, \cdot)$ under different sharpness parameters τ_w .

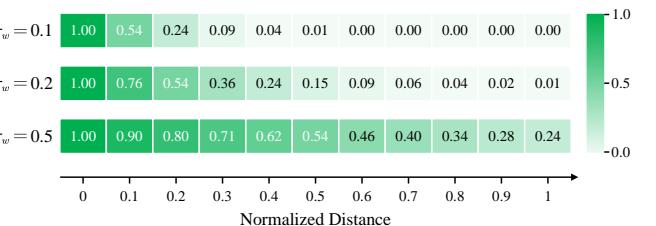


Figure 4: **Soft assignments $w(\cdot, \cdot)$ with different τ_w .** We use heat maps to visualize the soft assignments with different sharpness parameters τ_w . For clarity, we set the upper bound $\alpha = 1$ in the figure. Best viewed in color.

Leveraging the soft assignments for all sample pairs, we propose a soft contrastive loss to refine the traditional hard contrastive loss. Specifically, for a pair of projected embeddings (z_i, \tilde{z}_i) , we first calculate the softmax probability of

the relative similarity among all similarities as:

$$p(z_i, \tilde{z}_i) = \frac{\exp(\text{sim}(z_i, \tilde{z}_i)/\tau_c)}{\sum_{z' \in \mathcal{Z} \setminus \{z_i\}} \exp(\text{sim}(z_i, z')/\tau_c)}, \quad (4)$$

where $\text{sim}(\cdot, \cdot)$ refers to the cosine similarity, and τ_c is a temperature parameter used to adjust the scale. Based on $p(z_i, \tilde{z}_i)$, the soft contrastive loss is then defined as:

$$\begin{aligned} \mathcal{L}_{C,i} &= -\log p(z_i, \tilde{z}_i) \\ &- \sum_{\substack{x' \in \Omega \setminus \{x_i, \tilde{x}_i\} \\ z' \in \mathcal{Z} \setminus \{z_i, \tilde{z}_i\}}} w(x_i, x') \cdot \log p(z_i, z'), \end{aligned} \quad (5)$$

where $\Omega = \mathcal{X} \cup \tilde{\mathcal{X}}$ represents the union of the data spaces of the original and masked samples. By assigning soft weights to different sample pairs, the model is encouraged to better capture the inherent correlations across different samples. During the training process, the final soft contrastive loss \mathcal{L}_C is computed by summing and averaging $\mathcal{L}_{C,i}$ across all samples within a mini-batch. Notably, when $\forall w(x_i, x') = 0$, the soft contrastive loss reduces to the traditional hard contrastive loss.

Aggregate Reconstruction To further capture the fine-grained relationships between different samples, we incorporate an aggregator for weighted aggregation and reconstruction. Current approaches for masked EEG modeling typically reconstruct the masked portion based on the unmasked portion of a single masked sample (Lan et al. 2024; Pang et al. 2024), following the learning paradigm of MAE (He et al. 2022). However, this single-sample reconstruction strategy overlooks the interactions between samples, leading to a complex and inadequate reconstruction process.

To overcome this limitation, we introduce an aggregator that improves the traditional single-sample reconstruction process. Specifically, we first calculate the cosine similarity between each pair of projected embeddings (z_i, z_j) within a mini-batch, resulting in a similarity matrix \mathcal{S} . Based on the pairwise similarities in \mathcal{S} , the aggregator then performs weighted aggregation of the embedding h_i . The weighted aggregation process is defined as:

$$h_i^r = \sum_{z' \in \mathcal{Z} \setminus \{z_i\}} \left(\frac{\exp(\text{sim}(z_i, z')/\tau_c)}{\sum_{z'' \in \mathcal{Z} \setminus \{z_i\}} \exp(\text{sim}(z_i, z'')/\tau_c)} \cdot h'\right), \quad (6)$$

where $h' \in \mathcal{H} \setminus \{h_i\}$ represents the encoded embedding corresponding to the projected embedding z' , and \mathcal{H} denotes the embedding space of the encoder E . This approach allows for a more comprehensive reconstruction by aggregating complementary information and incorporating similar features from different samples during the reconstruction process, while suppressing interference from irrelevant noise samples. Finally, the reconstructed embedding h_i^r is fed into a lightweight decoder D to obtain the reconstructed sample x_i^r . Following the masked modeling paradigm, we use Mean Square Error (MSE) as the reconstruction loss for model optimization, which is defined as:

$$\mathcal{L}_{R,i} = \|x_i - x_i^r\|_2^2. \quad (7)$$

Similar to the soft contrastive loss \mathcal{L}_C , the final reconstruction loss \mathcal{L}_R is computed by summing and averaging $\mathcal{L}_{R,i}$ across all samples within a mini-batch.

The pre-training process of SCMM During the pre-training process, SCMM is trained by jointly optimizing \mathcal{L}_C and \mathcal{L}_R . The overall pre-training loss is defined as:

$$\mathcal{L} = \lambda_C \mathcal{L}_C + \lambda_R \mathcal{L}_R, \quad (8)$$

where λ_C and λ_R are trade-off hyperparameters, which are adaptively adjusted according to the homoscedastic uncertainty of each loss item. Algorithm 1 details the pre-training process of the proposed SCMM.

Algorithm 1: The pre-training process of SCMM.

Input: Unlabeled pre-training dataset $\mathcal{X} = \{x_i\}_{i=1}^N$. The number of pre-training epochs.

The pre-training process:

- 1: Randomly initialize the model parameters θ ;
 - 2: **for** epoch = 1 to epochs **do**
 - /*All operations are performed within a mini-batch*/
 - 3: Generate the masked sample \tilde{x}_i of each input EEG sample x_i using hybrid masking in **Eq. (1)**;
 - 4: Generate h_i and \tilde{h}_i by feeding x_i and \tilde{x}_i into E ;
 - 5: Generate z_i and \tilde{z}_i by feeding h_i and \tilde{h}_i into P ;
 - 6: Compute the normalized distance $D(x_i, x_j)$ for each pair of samples (x_i, x_j) using **Eq. (2)**;
 - 7: Generate the soft assignment $w(x_i, x_j)$ for each pair of samples (x_i, x_j) using **Eq. (3)**;
 - 8: Compute the soft contrastive loss \mathcal{L}_C using **Eq. (5)**;
 - 9: Compute the pairwise cosine similarity for each pair of projected embeddings (z_i, z_j) ;
 - 10: Generate the reconstructed embedding h_i^r of each h_i through weighted aggregation in **Eq. (6)**;
 - 11: Reconstruct x_i^r by feeding h_i^r into D ;
 - 12: Compute the reconstruction loss \mathcal{L}_R using **Eq. (7)**;
 - 13: Compute the pre-training loss \mathcal{L} using **Eq. (8)**;
 - 14: Update the model parameters θ ;
 - 15: **end for**
 - 16: **return** The pre-trained SCMM model f_θ .
-

Experiments

Datasets

We conduct extensive experiments on three publicly available datasets, SEED (Zheng and Lu 2015), SEED-IV (Zheng et al. 2018), and DEAP (Koelstra et al. 2011), to evaluate the model performance of SCMM. These datasets are diverse in terms of EEG equipment, emotional stimuli, data specifications, labeling approaches and subjects, making them well-suited for assessing the model's efficacy in cross-corpus EEG-based emotion recognition tasks. In our experiments, we use differential entropy (DE) features as inputs. Detailed descriptions of the datasets and pre-processing procedures are provided in Appendix A.

Methods	Same-Class		Different-Class	
	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³	SEED-IV ⁴ → SEED ³	SEED ³ → SEED-IV ⁴
BiDANN (Li et al. 2018)	49.24 / 10.49	60.46 / 11.17	-	-
TANN (Li et al. 2021)	58.41 / 07.16	60.75 / 10.61	-	-
PR-PL (Zhou et al. 2023)*	61.01 / 10.55	58.74 / 10.71	-	-
E ² STN (Zhou et al. 2024)	60.51 / 05.41	61.24 / 15.14	-	-
SimCLR (Tang et al. 2021)*	47.27 / 08.44	46.89 / 13.41	44.19 / 09.28	42.03 / 10.05
Mixup (Wickstrøm et al. 2022)*	56.86 / 16.83	55.70 / 16.28	54.55 / 17.95	45.79 / 15.16
MAE (He et al. 2022)*	86.49 / 10.57	<u>83.87 / 08.53</u>	<u>86.02 / 08.96</u>	<u>76.74 / 09.18</u>
JCFA (Liu et al. 2024a)*	67.53 / 12.36	62.40 / 07.54	65.99 / 14.04	52.67 / 05.86
SCMM (Ours)	91.61 / 07.56 (+05.12)	87.24 / 08.35 (+03.37)	91.26 / 07.91 (+05.24)	80.89 / 08.69 (+04.15)

Table 1: Experimental results on SEED and SEED-IV under two cross-corpus conditions: same-class and different-class. “*” indicates that the results are reproduced by ourselves. A → B denotes that A is the pre-training dataset, while B is the dataset for model fine-tuning and testing. **Best results** are highlighted in bold, while the second-best results are underlined.

Hard \mathcal{L}_C	Soft \mathcal{L}_C	\mathcal{L}_R	Same-Class		Different-Class	
			SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³	SEED-IV ⁴ → SEED ³	SEED ³ → SEED-IV ⁴
✓	✓	✓	90.08 / 09.24	84.43 / 11.83	90.26 / 08.71	77.79 / 08.04
		✓	90.73 / 08.48	85.07 / 11.05	90.96 / 08.36	78.32 / 07.19
		✓	89.68 / 09.32	84.24 / 11.90	89.45 / 09.10	77.24 / 09.14
✓	✓	✓	90.30 / 07.94	85.95 / 08.74	90.91 / 08.61	79.82 / 07.00
✓	✓	✓	91.61 / 07.56	87.24 / 08.35	91.26 / 07.91	80.89 / 08.69

Table 2: Ablation study on SEED and SEED-IV under two cross-corpus conditions: same-class and different-class.

Implementation Details

In the pre-training stage, we set r to 0.5 and μ to 0.1 for hybrid masking. We set α to 0.5, τ_w to 0.05, and τ_c to 0.5 for soft CL. The Adam optimizer is utilized with a learning rate of 5e-4 and a weight decay of 3e-4. The pre-training process is conducted over 200 epochs with a batch size of 256. We save the parameters θ from the final epoch as the pre-trained model. In the fine-tuning stage, we input the embeddings h_i generated by the encoder E into an emotion classifier consisting of a 2-layer MLP for final emotion recognition. The classifier is optimized using cross-entropy (CE) loss, and the fine-tuning process is conducted over 50 epochs with a batch size of 128. All experiments are conducted using Python 3.9 with PyTorch 1.13 on an NVIDIA GeForce RTX 3090 GPU. Further implementation details and hyperparameter analysis can be found in Appendix B and Appendix E.

Baselines and Experimental Setup

We compare the proposed SCMM against eight competitive baseline methods, including four conventional deep learning methods: BiDANN (Li et al. 2018), TANN (Li et al. 2021), PR-PL (Zhou et al. 2023), and E²STN (Zhou et al. 2024), as well as four self-supervised learning models: SimCLR (Chen et al. 2020; Tang et al. 2021), Mixup (Zhang et al. 2018; Wickstrøm et al. 2022), MAE (He et al. 2022), and JCFA (Liu et al. 2024a). Notably, E²STN and JCFA are specifically designed for cross-corpus EEG-based emo-

tion recognition. To ensure a fair comparison, we adopt a cross-corpus subject-independent protocol in our experiments, consistent with the setup used by JCFA. We evaluate the model performance using the average accuracy and standard deviation (ACC / STD %) across all subjects in the test set. More details about baseline methods and experimental settings are provided in Appendix C.

Experimental Results

Evaluation on the SEED and SEED-IV datasets To fully validate the model performance of SCMM, we conduct extensive experiments under two cross-corpus conditions: (1) same-class and (2) different-class.

(1) For the same-class cross-corpus validation, we conduct two experiments on the SEED and SEED-IV 3-category datasets: pre-training on SEED-IV and fine-tuning on SEED (SEED-IV³ → SEED³), and pre-training on SEED and fine-tuning on SEED-IV (SEED³ → SEED-IV³). In both experiments, all samples corresponding to fear emotions in the SEED-IV dataset are excluded. The left two columns in Table 1 present the comparison results, indicating that SCMM achieves SOTA performance in both experiments. Specifically, our model achieves classification accuracies of 91.61% and 87.24% with standard deviations of 7.56% and 8.35% in the SEED-IV³ → SEED³ and SEED³ → SEED-IV³ experiments, outperforming the second-best method MAE by accuracies of 5.12% and 3.37%, respectively.

Methods	DEAP → SEED ³	SEED ³ → DEAP (Valence)	SEED ³ → DEAP (Arousal)
SimCLR (Tang et al. 2021)*	53.12 / 13.12	53.75 / 03.61	51.79 / 04.54
Mixup (Wickstrøm et al. 2022)*	48.75 / 14.37	60.62 / 08.68	60.11 / 07.69
MAE (He et al. 2022)*	<u>83.69 / 10.10</u>	<u>72.19 / 07.24</u>	<u>70.50 / 06.30</u>
JCFA (Liu et al. 2024a)*	64.69 / 12.28	61.59 / 06.26	61.06 / 07.37
SCMM (Ours)	91.70 / 08.07 (+08.01)	73.96 / 06.75 (+01.77)	72.66 / 05.67 (+02.16)

Table 3: Experimental results for cross-corpus EEG-based emotion recognition on SEED and DEAP.

(2) For the different-class cross-corpus validation, we conduct another two experiments, denoted as SEED-IV⁴ → SEED³ and SEED³ → SEED-IV⁴. These experiments aim to evaluate the model performance when the pre-training and fine-tuning datasets contain different emotion categories. Experimental results in the right two columns of Table 1 demonstrate that SCMM achieves the best performance in both experiments. Specifically, our model achieves classification accuracies of 91.26% and 80.89% with standard deviations of 7.91% and 8.69% in the SEED-IV⁴ → SEED³ and SEED³ → SEED-IV⁴ experiments, surpassing the second-best method MAE by 5.24% and 4.15% in accuracies, respectively. Note that the traditional CL-based models SimCLR, Mixup, and JCFA exhibit relatively poor performance, primarily due to their use of raw EEG signals as inputs. In addition, SCMM significantly outperforms the traditional masked modeling framework MAE, highlighting the superiority of SCMM. In summary, extensive experimental results on the SEED and SEED-IV datasets confirm that our model exhibits superior performance in multiple cross-corpus EEG-based emotion recognition tasks.

Ablation study To assess the validity of each module in SCMM, we conduct a comprehensive ablation study on the SEED and SEED-IV datasets. Specifically, we design five different models below. (1) **Hard \mathcal{L}_C** trains the model with only hard contrastive loss. (2) **Soft \mathcal{L}_C** trains the model with only soft contrastive loss. (3) \mathcal{L}_R trains the model with only reconstruction loss. (4) **Hard \mathcal{L}_C and \mathcal{L}_R** trains the model using both hard contrastive loss and reconstruction loss. (5) **Soft \mathcal{L}_C and \mathcal{L}_R** trains the model using both soft contrastive loss and reconstruction loss, representing the full model.

Table 2 presents the results of ablation experiments. Specifically, (1) **Hard \mathcal{L}_C** and (2) **Soft \mathcal{L}_C** : the first two models compare the performance of hard CL and soft CL. Experimental results indicate that our well-designed soft contrastive learning effectively improves the classification accuracy. (3) \mathcal{L}_R : the third model removes the soft contrastive loss and trains the model with only reconstruction loss. The results show that the model performs worst in the absence of the contrastive learning constraint. (4) **Hard \mathcal{L}_C and \mathcal{L}_R** and (5) **Soft \mathcal{L}_C and \mathcal{L}_R** : the last two models compare the performance when combining hard and soft contrastive loss with reconstruction loss, respectively. Experimental results demonstrate that the full model achieves the best performance in all experiments, indicating that SCMM significantly enhances the model performance and stability by combining soft contrastive learning and aggregate recon-

struction. This improvement is evident under different cross-corpus conditions, demonstrating the feasibility of extending SCMM to real-life aBCI applications.

Discussion

Generalization capability analysis To further validate the generalization capability of SCMM, we conduct additional experiments on the SEED and DEAP datasets under a different-class cross-corpus scenario, denoted as DEAP → SEED³, SEED³ → DEAP (Valence) and SEED³ → DEAP (Arousal). It is notable that the EEG acquisition equipment, emotional stimuli, data specifications, labeling approaches and subjects are completely different between the two datasets. Table 3 presents the experimental results of SCMM compared to existing methods. Specifically, for the DEAP → SEED³ experiment, SCMM achieves an accuracy of 91.70% with a standard deviation of 8.07%, outperforming the second-best method MAE by an accuracy of 8.01%. For the SEED³ → DEAP (Valence) and SEED³ → DEAP (Arousal) experiments, our model achieves classification accuracies of 73.96% and 72.66% with standard deviations of 6.75% and 5.67%, surpassing the second-best method MAE by 1.77% and 2.16% in accuracies. Experimental results demonstrate that the proposed SCMM maintains excellent performance even when the pre-training and fine-tuning datasets are completely different, highlighting its superior generalization capability.

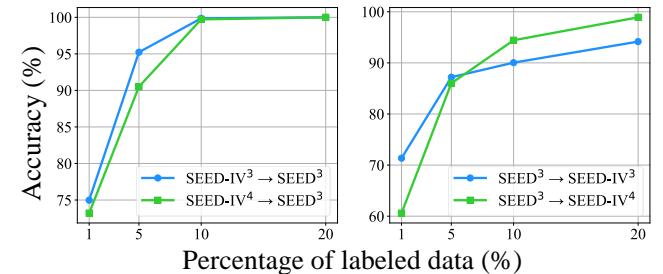


Figure 5: Model performance with limited labeled data for fine-tuning on SEED and SEED-IV.

Model performance with limited fine-tuning data We further explore the model performance of SCMM on the SEED and SEED-IV datasets when fine-tuning labeled data is limited. Specifically, we randomly select 1%, 5%, 10% and 20% of labeled samples from the fine-tuning dataset for

Method	ACC / STD (%)			
	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³	SEED-IV ⁴ → SEED ³	SEED ³ → SEED-IV ⁴
OS	91.61 / 07.56	87.24 / 08.35	91.25 / 07.91	80.89 / 08.69
ES	89.99 / 10.25	85.75 / 14.00	90.31 / 08.59	79.04 / 06.95
	DEAP → SEED ³	SEED ³ → DEAP (Valence)	SEED ³ → DEAP (Arousal)	-
OS	91.70 / 08.01	73.96 / 06.75	72.66 / 05.67	-
ES	90.64 / 07.97	72.75 / 07.06	71.58 / 05.72	-

Table 4: Comparison of soft contrastive learning in the original data space (OS) and embedding space (ES).

Strategies	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³
Random	90.30 / 08.80	84.63 / 10.99
Channel	90.25 / 08.68	85.91 / 10.97
Hybrid	91.61 / 07.56	87.24 / 08.35

Table 5: Comparative experiments of different masking strategies on the SEED and SEED-IV 3-category datasets.

model fine-tuning, while the remaining samples are used for testing. Figure 5 shows the classification accuracy curves. For the SEED-IV³ → SEED³ and SEED-IV⁴ → SEED³ experiments, SCMM achieves classification accuracies exceeding 70% with only 1% of labeled data. The accuracies significantly improve as the proportion of labeled samples increases, reaching close to 100% with 10% of labeled data. Meanwhile, our model achieves classification accuracies over 60% and 70% with only 1% of labeled data in the SEED³ → SEED-IV³ and SEED³ → SEED-IV⁴ experiments. The classification accuracies exceed 90% when fine-tuning with 10% of labeled samples in both experiments. In summary, experimental results indicate that SCMM maintains superior performance even with limited labeled data for fine-tuning, showing its outstanding robustness and potential in few-shot scenarios. The complete results and computational complexity analysis are provided in Appendix F.

Comparison of different masking strategies To assess the impact of various masking strategies, we conduct comparative experiments on the SEED and SEED-IV 3-category datasets using three strategies: random, channel, and hybrid masking. Table 5 presents the experimental results, showing that the hybrid masking strategy achieves the highest accuracy and the lowest standard deviation in both experiments. This suggests that the integration of different masking approaches significantly improves the model performance and stability. More details on the different masking strategies are presented in Appendix D.

Comparison of soft CL in the original data space and embedding space While soft contrastive learning has been explored across various domains, most methods focus on computing soft assignments for contrastive loss in the embedding space (Dwibedi et al. 2021; Yèche et al. 2021). However, we argue that utilizing similarities in the original

data space offers superior self-supervision and is particularly well-suited for EEG emotional data. To validate this, we conduct additional experiments on the SEED, SEED-IV and DEAP datasets to verify the effectiveness of soft contrastive learning in the original data space. Specifically, we modify the metric function $Dist(\cdot, \cdot)$ to use similarities between projected embeddings, shifting the computation of soft assignments from the original data space to the embedding space. Table 4 presents the experimental results, demonstrating that soft contrastive learning in the original data space consistently outperforms the embedding space in all experiments. Furthermore, this approach allows for pre-computing cosine similarities between original sample pairs offline, thus reducing computational resource requirements and improving training efficiency.

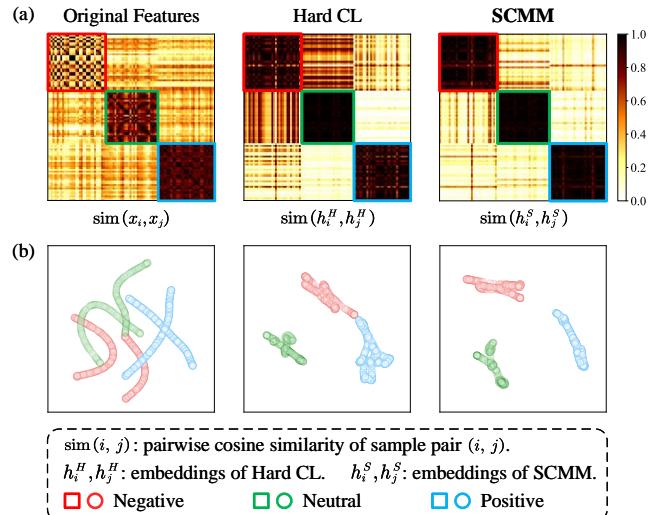


Figure 6: (a) Heat maps of pairwise similarity matrices. (b) t-SNE visualization of the learned embeddings. Best viewed in color. Zoom in for better view.

Visualization To evaluate whether sample-wise relationships are preserved in the encoder, we randomly select 100 test samples from the SEED dataset, and visualize the pairwise cosine similarity between sample pairs. Additionally, we select all test samples of one subject from the SEED

dataset and visualize the learned embeddings of SCMM using t-SNE (Van der Maaten and Hinton 2008). Figure 6(a) presents heat maps of pairwise similarity matrices, where darker colors indicate higher similarity between samples. Traditional hard CL identifies only coarse-grained relationships across samples from different emotion categories, especially for the most challenging-to-recognize negative and neutral emotions. In contrast, SCMM effectively captures the fine-grained relationships between samples of different categories. Moreover, the results of t-SNE visualization in Fig. 6(b) indicate that the proposed SCMM better clusters samples within the same category and increases the inter-class distance compared to traditional hard CL, thus enhancing the classification performance. More visualization results can be found in Appendix G.

Conclusion

This paper proposes a novel self-supervised pre-training framework, **Soft Contrastive Masked Modeling (SCMM)**, for cross-corpus EEG-based emotion recognition. Unlike traditional contrastive learning models, SCMM integrates soft contrastive learning with a hybrid masking strategy to effectively capture the "short-term continuity" characteristics inherent in human emotions, and produce stable and generalizable EEG representations. Additionally, an aggregator is developed to weightedly aggregate complementary information from multiple close samples, thereby enhancing fine-grained feature representation capability in the modeling process. Extensive experiments on three well-recognized datasets show that SCMM consistently achieves SOTA performance in cross-corpus EEG-based emotion recognition tasks under both same-class and different-class conditions. Comprehensive ablation study and parameter analysis confirm the superior performance and robustness of SCMM. Visualization results indicate that our model effectively reduces the distance between similar samples within the same category, and captures more fine-grained relationships across samples. These findings suggest that SCMM enhances the feasibility of extending the proposed method to real-life aBCI applications.

References

- Canal, F. Z.; Müller, T. R.; Matias, J. C.; Scotton, G. G.; de Sa Junior, A. R.; Pozzebon, E.; and Sobieranski, A. C. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582: 593–617.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.
- Chien, W.-S.; Yang, H.-C.; and Lee, C.-C. 2020. Cross corpus physiological-based emotion recognition using a learnable visual semantic graph convolutional network. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2999–3006.
- Davidson, R. J. 1998. Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & emotion*, 12(3): 307–330.
- Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9588–9597.
- Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2352–2359.
- Gao, P.; Liu, T.; Liu, J.-W.; Lu, B.-L.; and Zheng, W.-L. 2024. Multimodal Multi-View Spectral-Spatial-Temporal Masked Autoencoder for Self-Supervised Emotion Recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1926–1930. IEEE.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.
- He, Z.; Zhong, Y.; and Pan, J. 2022. An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition. *Computers in Biology and Medicine*, 141: 105048.
- Houben, M.; Van Den Noortgate, W.; and Kuppens, P. 2015. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological bulletin*, 141(4): 901.
- Hu, X.; Chen, J.; Wang, F.; and Zhang, D. 2019. Ten challenges for EEG-based affective computing. *Brain Science Advances*, 5(1): 1–20.
- Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1): 18–31.
- Lan, Y.-T.; Jiang, W.-B.; Zheng, W.-L.; and Lu, B.-L. 2024. CEMOAE: A Dynamic Autoencoder with Masked Channel Modeling for Robust EEG-Based Emotion Recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1871–1875. IEEE.
- Li, R.; Wang, Y.; Zheng, W.-L.; and Lu, B.-L. 2022. A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6–14.
- Li, Y.; Fu, B.; Li, F.; Shi, G.; and Zheng, W. 2021. A novel transferability attention neural network model for EEG emotion recognition. *Neurocomputing*, 447: 92–101.
- Li, Y.; Zheng, W.; Cui, Z.; Zhang, T.; and Zong, Y. 2018. A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition. In *IJCAI*, 1561–1567.

- Liu, Q.; Zhou, Z.; Wang, J.; and Liang, Z. 2024a. Joint Contrastive Learning with Feature Alignment for Cross-Corpus EEG-based Emotion Recognition. arXiv:2404.09559.
- Liu, Z.; Alavi, A.; Li, M.; and Zhang, X. 2024b. Guidelines for Augmentation Selection in Contrastive Learning for Time Series Classification. arXiv:2407.09336.
- Noroozi, F.; Corneanu, C. A.; Kamińska, D.; Sapiński, T.; Escalera, S.; and Anbarjafari, G. 2018. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 12(2): 505–523.
- Pang, M.; Wang, H.; Huang, J.; Vong, C.-M.; Zeng, Z.; and Chen, C. 2024. Multi-Scale Masked Autoencoders for Cross-Session Emotion Recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.
- Rayatdoost, S.; and Soleymani, M. 2018. Cross-corpus EEG-based emotion recognition. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.
- Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.
- Ryumina, E.; Dresvyanskiy, D.; and Karpov, A. 2022. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 514: 435–450.
- Schuller, B.; Vlasenko, B.; Eyben, F.; Wöllmer, M.; Stuhlsatz, A.; Wendemuth, A.; and Rigoll, G. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2): 119–131.
- Shen, X.; Liu, X.; Hu, X.; Zhang, D.; and Song, S. 2022. Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14(3): 2496–2511.
- Singh, Y. B.; and Goel, S. 2022. A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 492: 245–263.
- Tang, C. I.; Perez-Pozuelo, I.; Spathis, D.; and Mascolo, C. 2021. Exploring Contrastive Learning in Human Activity Recognition for Healthcare. arXiv:2011.11542.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).
- Wang, Y.; Zhang, B.; and Tang, Y. 2024. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 628–636.
- Wickstrøm, K.; Kampffmeyer, M.; Mikalsen, K. Ø.; and Jenssen, R. 2022. Mixing up contrastive learning: Self-supervised representation learning for time series. *Pattern Recognition Letters*, 155: 54–61.
- Yèche, H.; Dresdner, G.; Locatello, F.; Hüser, M.; and Rätsch, G. 2021. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, 11964–11974. PMLR.
- Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8980–8987.
- Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. arXiv:1710.09412.
- Zhang, K.; Wen, Q.; Zhang, C.; Cai, R.; Jin, M.; Liu, Y.; Zhang, J. Y.; Liang, Y.; Pang, G.; Song, D.; et al. 2024. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhang, X.; Zhao, Z.; Tsiligkaridis, T.; and Zitnik, M. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35: 3988–4003.
- Zhang, Z.; Liu, Y.; and Zhong, S.-h. 2022. GANSER: A self-supervised data augmentation framework for EEG-based emotion recognition. *IEEE Transactions on Affective Computing*, 14(3): 2048–2063.
- Zhang, Z.; Weninger, F.; Wöllmer, M.; and Schuller, B. 2011. Unsupervised learning in cross-corpus acoustic emotion recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 523–528. IEEE.
- Zhao, L.-M.; Yan, X.; and Lu, B.-L. 2021. Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 863–870.
- Zheng, W.-L.; Liu, W.; Lu, Y.; Lu, B.-L.; and Cichocki, A. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49(3): 1110–1122.
- Zheng, W.-L.; and Lu, B.-L. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3): 162–175.
- Zhong, P.; Wang, D.; and Miao, C. 2020. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3): 1290–1301.
- Zhou, R.; Zhang, Z.; Fu, H.; Zhang, L.; Li, L.; Huang, G.; Li, F.; Yang, X.; Dong, Y.; Zhang, Y.-T.; et al. 2023. PR-PL: A novel prototypical representation based pairwise learning framework for emotion recognition using EEG signals. *IEEE Transactions on Affective Computing*, 15(2): 657–670.
- Zhou, Y.; Li, F.; Li, Y.; Ji, Y.; Zhang, L.; and Chen, Y. 2024. Enhancing Cross-Dataset EEG Emotion Recognition: A Novel Approach with Emotional EEG Style Transfer Network. arXiv:2403.16540.

Appendix A Datasets

A.1 Dataset Description

We conduct extensive experiments on three well-recognized datasets, SEED (Zheng and Lu 2015), SEED-IV (Zheng et al. 2018), and DEAP (Koelstra et al. 2011), to evaluate the model performance of SCMM in cross-corpus EEG-based emotion recognition tasks. Table 6 provides a detailed description of the three datasets.

(1) **SEED** (Zheng and Lu 2015) was developed by Shanghai Jiao Tong University. The dataset used a 62-channel ESI NeuroScan System based on the international 10-20 system to record EEG signals from 15 subjects (7 males and 8 females) under different video stimuli at a sampling rate of 1kHz. Each subject participated in 3 sessions. In each session, each subject was required to watch 15 movie clips consisting of 3 different emotional states: negative, neutral and positive. Each emotional state contains a total of 5 movie clips, corresponding to 5 trials.

(2) **SEED-IV** (Zheng et al. 2018) used the same EEG acquisition equipment as the SEED dataset, but with different video stimuli, emotion categories and subjects. The dataset recorded EEG signals from 15 subjects under different video stimuli at a sampling rate of 1kHz. Each subject participated in 3 sessions. In each session, each subject was required to watch 24 movie clips containing 4 different emotions: sad, neutral, fear and happy. Each emotion contains a total of 6 movie clips, corresponding to 6 trials.

(3) **DEAP** (Koelstra et al. 2011) was constructed by Queen Mary University of London. The dataset used a 128-channel Biosemi ActiveTwo System to record EEG signals from specific 32 channels of 32 subjects (16 males and 16 females) while watching 40 one-minute music videos at a sampling rate of 512Hz. The 40 one-minute videos elicited different emotions according to the valence-arousal dimension. Specifically, the valence-arousal emotional model, first proposed by Russell (Russell 1980), places each emotional state on a two-dimensional scale. The first dimension represents valence, ranging from negative to positive, and the second dimension represents arousal, ranging from calm to exciting. Participants rated valence and arousal using a continuous scale of 1 to 9 after watching each video clip.

A.2 Pre-processing Procedures

For the SEED and SEED-IV datasets, the raw EEG signals were initially downsampled to 200Hz and filtered through a bandpass filter of 0.3-50Hz to filter noise and remove artifacts. Then, the data were divided into multiple non-overlapping segments using sliding windows of 1 second for SEED and 4 seconds for SEED-IV, respectively. After that, we extracted differential entropy (DE) features for each channel of each segment at five frequency bands: Delta (1-4Hz), Theta (4-8Hz), Alpha (8-14Hz), Beta (14-31Hz), and Gamma (31-50Hz). Finally, the DE features from 62 channels and 5 bands were formed into a feature matrix of shape 62×5 , which serves as input to the SCMM model. The extraction of DE features can be expressed as:

$$DE(x) = \frac{1}{2} \log(2\pi e \sigma^2), \quad (9)$$

Here, x represents an EEG signal segment of a specific length that approximately obeys a Gaussian distribution $N(\mu, \sigma^2)$, where σ denotes the standard deviation of x , and e is the Euler constant.

For the DEAP dataset, the raw EEG signals were first downsampled to 128Hz and denoised by a bandpass filter of 4-45Hz. Subsequently, the data were segmented into multiple non-overlapping segments using a sliding window of 1s. Similar to the SEED and SEED-IV datasets, DE features were extracted for each channel of each segment at five frequency bands. Finally, the DE features from 32 channels and 5 bands were formed into a feature matrix of shape 32×5 as input to the model. During the experiments, we divided the continuous labels using a fixed threshold of 5 to convert them to binary classification tasks (low / high).

A.3 Handling Different Number of Channels

Since the SEED-series datasets and the DEAP dataset contain different numbers of electrodes (channels), we require channel processing before inputting DE features into the model. Specifically, we consider the fine-tuning dataset as an anchor. When the number of channels in the fine-tuning dataset is less than in the pre-training dataset, we select data from the corresponding channels in the pre-training dataset and drop the data from the redundant channels as inputs (e.g., pre-training on SEED and fine-tuning on DEAP). Conversely, when the number of channels in the fine-tuning dataset is greater than in the pre-training dataset, we fill the missing channel data with zeros in the pre-training dataset to match the fine-tuning dataset (e.g., pre-training on DEAP and fine-tuning on SEED).

Appendix B Implementation Details

To reduce computational load while maintaining model performance, we adopt a lightweight design for each module of SCMM. Specifically, we use a 3-layer 1D CNN for the encoder E and a 2-layer MLP for the projector P . For the lightweight decoder D , we utilize a single-layer MLP for reconstruction. Regarding hyperparameter selection in the pre-training stage, we set r to 0.5 and μ to 0.1 for hybrid masking, i.e., the percentage of random masking and channel masking is 0.9 and 0.1, respectively. We use the negative of cosine similarity as the metric function $Dist(\cdot, \cdot)$, and we set α to 0.5, τ_w to 0.05 and τ_c to 0.5 for soft contrastive learning. We use Adam as the optimizer with a learning rate of 5×10^{-4} and an L2-norm penalty coefficient 3×10^{-4} . The pre-training process is conducted over 200 epochs with a batch size of 256. We save the model parameters θ from the final epoch as the pre-trained SCMM. In the fine-tuning stage, we input the encoded embeddings h_i into a classifier consisting of a 2-layer fully connected network for final emotion recognition. The Adam optimizer is utilized with a learning rate of 5×10^{-4} and a weight decay of 3×10^{-4} . The number of fine-tuning epochs is set to 50 for the SEED and SEED-IV datasets and 500 for the DEAP dataset, with a batch size of 128. For efficient deployment and testing of the model, the pre-trained SCMM is optimized solely using cross-entropy loss during fine-tuning. All experiments

Datasets	Subjects	Sessions × Trials	Channels	Sampling Rate	Classes
SEED	15	3×15	62	1kHz	3 (Negative, Neutral, Positive)
SEED-IV	15	3×15	62	1kHz	4 (Sad, Neutral, Fear, Happy)
DEAP	32	1×40	32	512Hz	Valence: 1 - 9, Arousal: 1 - 9

Table 6: Detailed description of the experimental datasets.

	Pre-training	Fine-tuning
Encoder	3-layer 1D CNN	
Projector	2-layer MLP	
Decoder	single-layer MLP	
Classifier	-	2-layer MLP
Masking ratio r	0.5	-
Threshold μ	0.1	-
Upper bound α	0.5	-
Sharpness τ_w	0.05	-
Temperature τ_c	0.5	-
Epoch	200	50, 500
Optimizer	Adam	
Learning rate	5×10^{-4}	
Weight decay	3×10^{-4}	
Batch size	256	128

Table 7: Hyperparameter settings of SCMM.

are conducted using Python 3.9 with PyTorch 1.13 on an NVIDIA GeForce 3090 GPU.

Appendix C Baseline Methods and Experimental Settings

C.1 Baseline Methods

We compare the proposed SCMM against eight competitive methods, including four traditional deep learning methods: BiDANN (Li et al. 2018), TANN (Li et al. 2021), PR-PL (Zhou et al. 2023) and E²STN (Zhou et al. 2024), as well as four self-supervised learning models: SimCLR (Chen et al. 2020; Tang et al. 2021), Mixup (Zhang et al. 2018; Wickstrøm et al. 2022), MAE (He et al. 2022) and JCFA (Liu et al. 2024a). Details of the eight baseline methods are summarized as follows.

- **BiDANN** (Li et al. 2018): The bi-hemispheres domain adversarial neural network mapped the EEG data of both left and right hemispheres into discriminative feature spaces separately to address domain shifts in EEG-based emotion recognition tasks.

- **TANN** (Li et al. 2021): The transferable attention neutral network is a novel transfer learning methods which learned the discriminative information from EEG signals using local and global attention mechanisms.
- **PR-PL** (Zhou et al. 2023): The prototypical representation based pairwise learning framework adopted pairwise learning to model the relative relationships between EEG sample pairs in terms of prototypical representations, addressing the critical issues of individual differences and noise labels in cross-subject scenarios.
- **E²STN** (Zhou et al. 2024): The emotional EEG style transfer network integrated content information from the source domain with style information from the target domain, achieving superior performance in cross-corpus EEG-based emotion recognition tasks.
- **SimCLR** (Chen et al. 2020; Tang et al. 2021): A seminal work in self-supervised contrastive learning, first proposed for computer vision, has been extended to human activity recognition.
- **Mixup** (Zhang et al. 2018; Wickstrøm et al. 2022): A novel CL-based data augmentation method that aimed to correctly predict the mixing proportion of two samples, has been applied to time series analysis.
- **MAE** (He et al. 2022): A groundbreaking work in the field of mask modeling, which proposed to reconstruct the masked portion based on the unmasked portion of the masked sample, has achieved remarkable success in a wide range of fields.
- **JCFA** (Liu et al. 2024a): The joint contrastive learning framework performed joint contrastive learning across two domains to synchronize the time- and frequency-based embeddings of the same EEG sample in the latent time-frequency space, achieving state-of-the-art performance in cross-corpus scenarios.

To ensure a fair comparison, we adopt the same encoder, projector, decoder, and classifier structures for SimCLR, Mixup and MAE as used in SCMM. We use the default hyperparameters reported in the original paper for all methods in our experiments, unless otherwise specified. Additionally, for BiDANN, TANN, PR-PL, E²STN, MAE, and SCMM, the input samples are preprocessed 1-s DE features. In contrast, SimCLR, Mixup, and JCFA use preprocessed 1-s EEG signals as inputs, in accordance with the specific design of each model.

C.2 Experimental Settings

We adopt a cross-corpus subject-independent experimental protocol in the experiments, following the setup used in

JCFA. Specifically, samples from one dataset are used for pre-training, while samples of each subject from another dataset are used individually for fine-tuning and testing. During the fine-tuning process, we employ a leave-trials-out setting, where samples from a part of trials in each session of each subject in the fine-tuning dataset are used for fine-tuning, and the remaining trials are used for testing. For example, SCMM is pre-trained on the SEED-IV dataset and fine-tuned on the SEED dataset. The allocation of samples for fine-tuning/testing is as follows: 9/6 for the SEED dataset, 12/6 for the SEED-IV 3-category dataset, 16/8 for the SEED-IV 4-category dataset, and 24/16 for the DEAP dataset. This approach effectively avoids information leakage. The detailed experimental settings and data division for pre-training and fine-tuning are summarized in Table 8.

Appendix D Masking Strategy

This paper introduces a novel hybrid masking strategy to generate diverse masked samples by considering both channel and feature relationships. To compare our approach with traditional masking strategies, we explore three different masking rules: random masking, channel masking, and hybrid masking. Figure 3 illustrates the difference between these three masking strategies.

(1) **Random Masking:** Generate masks using a binomial distribution to randomly mask samples in the feature dimension, setting the values of the masked features to zero.

(2) **Channel Masking:** Generate masks using a binomial distribution to randomly mask samples in the channel dimension, setting the values of all features within the masked channels to zero.

(3) **Hybrid Masking:** Generate a probability matrix using a uniform distribution that proportionally mixes masks generated from random masking and channel masking.

Our proposed hybrid masking strategy is highly flexible and can be extended to various datasets by integrating multiple masking strategies in different ratios, which is exceptionally suitable for data with rich semantic information. This approach effectively generates more diverse masked samples, encouraging the model to comprehensively capture the inherent relationships of the data.

Appendix E Hyperparameter Analysis

We conduct comprehensive experiments to verify the hyperparameter sensitivity of SCMM on the SEED and SEED-IV 3-category datasets. The hyperparameters examined include the masking ratio r , probability threshold μ , metric function $Dist(\cdot, \cdot)$, upper bound α , sharpness τ_w , and temperature τ_c . The complete experimental results are presented in the following sections.

E.1 Masking Ratio

Table 9 shows the model performance in the $SEED-IV^3 \rightarrow SEED^3$ and $SEED^3 \rightarrow SEED-IV^3$ experiments when using different masking ratios (0.1, 0.25, 0.5, and 0.75). Experimental results demonstrate that the proposed SCMM consistently achieves the best performance when the masking ratio is set to $r = 0.5$.

E.2 Probability Threshold

Table 10 presents the model performance using different probability thresholds (ranging from 0 to 1) for hybrid masking. Here, $\mu = 0$ refers to the exclusive use of the random masking strategy, while $\mu = 1$ indicates the exclusive use of the channel masking strategy. Experimental results show that SCMM achieves the best performance in the $SEED-IV^3 \rightarrow SEED^3$ and $SEED^3 \rightarrow SEED-IV^3$ experiments when the probability threshold is set to $\mu = 0.1$ (i.e., the ratio of random masking and channel masking is 9:1).

E.3 Metric Function

We evaluate the model performance of SCMM using different metric functions in soft assignments $w(\cdot, \cdot)$. Experimental results in Table 11 indicate that SCMM performs best on the SEED and SEED-IV 3-category datasets when cosine similarity is used as the metric function.

E.4 Upper Bound

We assess the impact of different upper bounds on calculating soft assignments $w(\cdot, \cdot)$. To restrict the soft assignments to the range of 0 to 1, we explore the model performance of SCMM with upper bounds of 0.25, 0.5, 0.75, and 1. Experimental results in Table 12 illustrate that the proposed SCMM achieves the best performance in the $SEED-IV^3 \rightarrow SEED^3$ experiment when the upper bound is set to $\alpha = 0.5$.

E.5 Sharpness

Table 13 explores the impact of different sharpness parameters in the $SEED-IV^3 \rightarrow SEED^3$ experiment, as shown in Table 13. Experimental results indicate that our model achieves the best performance with a sharpness parameter of $\tau_w = 0.05$.

E.6 Temperature

We conduct comparison experiments on the SEED and SEED-IV 3-category datasets to explore the impact of temperature parameter on model performance. Experimental results in Table 14 show that SCMM consistently achieves the best performance in both experiments when the temperature parameter is set to $\tau_c = 0.5$.

F Model Performance and Computational Complexity Analysis of SCMM

To investigate the trade-off between model performance and computational complexity in few-shot scenarios, we assess the classification accuracy, time cost, and the number of trainable parameters of SCMM when fine-tuning with limited labeled data. Experimental results in Table 15 show that the model performance of SCMM significantly improves as the number of fine-tuning labeled samples increases. Meanwhile, the inference time cost remains low due to our lightweight design. Additionally, our model achieves superior classification performance in cross-corpus EEG-based emotion recognition tasks with very few parameters. In summary, the proposed SCMM effectively balances model performance and computational complexity.

Evaluations	Scenarios	Pre-training	Fine-tuning / Testing
Same-Class	SEED-IV ³ → SEED ³	SEED-IV, 3-category	SEED: 9 / 6 trials in each session per subject
	SEED ³ → SEED-IV ³	SEED	SEED-IV, 3-category: 12 / 6 trials in each session per subject
Different-Class	SEED-IV ⁴ → SEED ³	SEED-IV, 4-category	SEED: 9 / 6 trials in each session per subject
	SEED ³ → SEED-IV ⁴	SEED	SEED-IV, 4-category: 16 / 8 trials in each session per subject
Different-Class	DEAP → SEED ³	DEAP	SEED: 9 / 6 trials in each session per subject
	SEED ³ → DEAP (Valence)	SEED	DEAP (Valence): 24 / 16 trials per subject
	SEED ³ → DEAP (Arousal)	SEED	DEAP (Arousal): 24 / 16 trials per subject

Table 8: Pre-training and fine-tuning scenarios under two conditions for cross-corpus EEG-based emotion recognition.

Ratio	ACC / STD (%)	
	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³
0.1	91.19 / 08.07	85.05 / 11.68
0.25	91.50 / 07.69	85.28 / 09.91
0.5	91.61 / 07.56	87.24 / 08.35
0.75	90.24 / 08.08	83.49 / 13.19

Table 9: Hyperparameter sensitivity analysis of masking ratio r on SEED and SEED-IV.

G Visualization

G.1 Intra- and Inter-Class Similarities

To assess the quality of the embeddings learned by SCMM, we randomly select one subject from the SEED dataset, and calculate both the average intra- and inter-class cosine similarities between the learned embeddings of all test samples, as shown in Fig. 7. It is evident that the proposed SCMM produces embeddings with higher intra-class similarity compared to traditional hard contrastive learning. In addition, the average inter-class similarity of the embeddings learned by SCMM is significantly lower than that of hard CL. In summary, visualization results confirm that the soft contrastive learning strategy designed in SCMM effectively clusters samples within the same category while distinctly separating samples from different categories, thus enhancing the model’s discriminative capabilities.

G.2 Reconstruction Quality

To verify the effectiveness of the aggregator designed in SCMM, we compare the reconstruction quality of the single-sample reconstruction paradigm (MAE) with the aggregate reconstruction paradigm (SCMM) on the DEAP dataset. For clarity, we flatten both the original input sample and the reconstructed sample into one-dimensional vectors with dimensions $C \times F$ (channels \times features). The results depicted in Fig. 8 illustrate that our model achieves lower reconstruction loss (MSE) and better sample reconstruction.

Threshold	ACC / STD (%)	
	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³
0 (Random)	90.30 / 08.80	84.63 / 10.99
0.1	91.61 / 07.56	87.24 / 08.35
0.2	90.93 / 08.18	86.51 / 08.44
0.3	90.84 / 07.70	86.66 / 09.98
0.4	89.34 / 08.70	86.94 / 09.81
0.5	91.00 / 08.22	87.16 / 11.09
0.6	89.93 / 08.90	85.83 / 09.65
0.7	91.08 / 07.64	85.14 / 11.60
0.8	89.52 / 08.29	86.10 / 12.48
0.9	89.42 / 08.18	86.05 / 11.75
1 (Channel)	90.25 / 08.68	85.91 / 10.97

Table 10: Hyperparameter sensitivity analysis of threshold μ on SEED and SEED-IV.

Metrics	ACC / STD (%)	
	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³
Manhattan	90.09 / 09.06	85.83 / 11.14
Euclidean	90.86 / 08.57	85.40 / 13.51
Cosine	91.61 / 07.56	87.24 / 08.35

Table 11: Hyperparameter sensitivity analysis of metric function $Dist(\cdot, \cdot)$ on SEED and SEED-IV.

Upper Bound	ACC / STD (%)	
	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³
0.25	91.21 / 08.54	
0.5	91.61 / 07.56	
0.75	91.12 / 08.25	
1	90.50 / 08.00	

Table 12: Hyperparameter sensitivity analysis of upper bound α in the SEED-IV³ → SEED³ experiment.

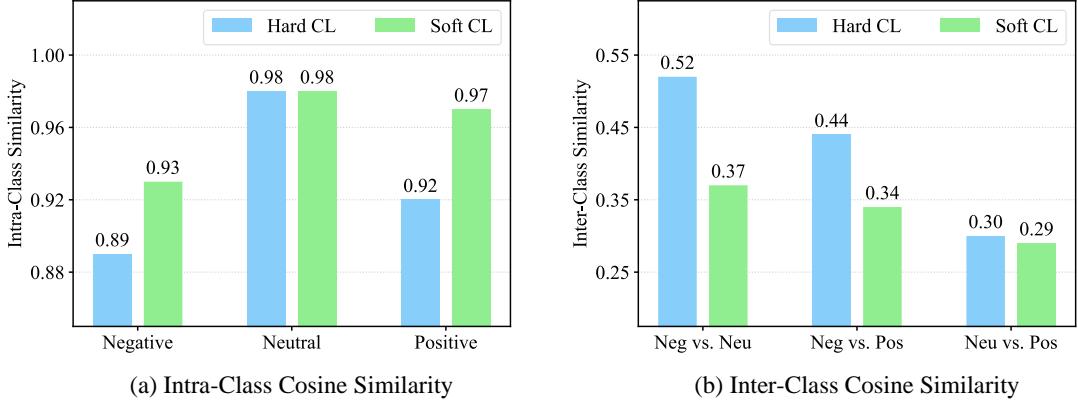


Figure 7: **Intra- and inter-class cosine similarities of embeddings learned by hard CL and soft CL.** (a) is the average intra-class cosine similarity, and (b) is the average inter-class cosine similarity. Best viewed in color.

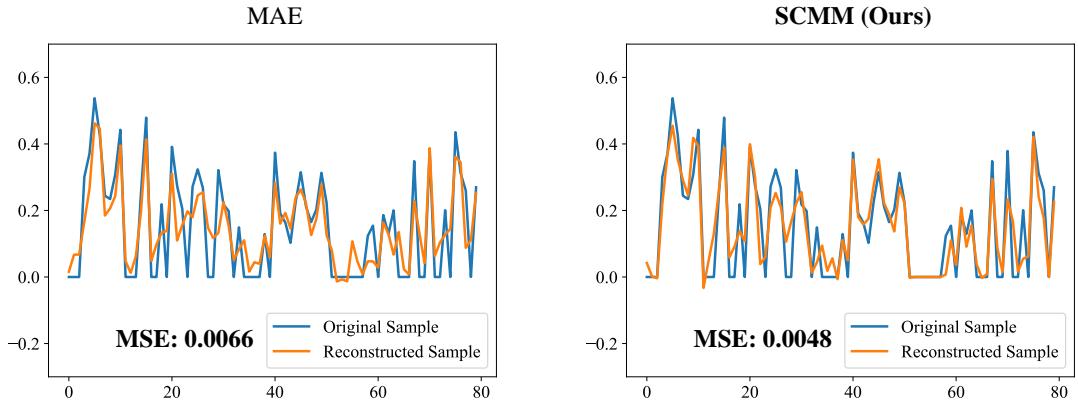


Figure 8: **Comparison of reconstruction quality.** We visualize the reconstruction results to compare the single-sample reconstruction paradigm (MAE) with the aggregate reconstruction paradigm (SCMM). This case is from the DEAP dataset. Best viewed in color.

Sharpness	ACC / STD (%)
	SEED-IV ³ → SEED ³
0.01	90.93 / 09.13
0.05	91.61 / 07.56
0.1	91.08 / 09.00
0.2	90.59 / 07.40
0.5	90.63 / 09.16
1	90.09 / 08.62
2	89.58 / 09.26

Table 13: Hyperparameter sensitivity analysis of sharpness τ_w in the SEED-IV³ → SEED³ experiment.

Temperature	ACC / STD (%)	
	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³
0.05	90.97 / 08.17	85.95 / 11.14
0.2	91.16 / 07.85	86.82 / 11.46
0.5	91.61 / 07.56	87.24 / 08.35
2	90.62 / 08.66	86.73 / 10.18

Table 14: Hyperparameter sensitivity analysis of temperature τ_c on SEED and SEED-IV.

Metrics	1% of labeled data			
	SEED-IV ³ → SEED ³	SEED ³ → SEED-IV ³	SEED-IV ⁴ → SEED ³	SEED ³ → SEED-IV ⁴
ACC / STD (%)	74.98 / 17.00	71.35 / 16.29	73.20 / 16.09	60.59 / 23.72
Time Cost (s)	232.35	97.11	232.35	130.47
5% of labeled data				
ACC / STD (%)	95.23 / 06.98	87.20 / 16.12	90.51 / 12.43	86.01 / 09.76
Time Cost (s)	240.48	100.23	240.48	132.90
10% of labeled data				
ACC / STD (%)	99.86 / 00.31	90.04 / 16.94	99.72 / 00.80	94.43 / 05.60
Time Cost (s)	250.81	102.03	250.81	135.67
20% of labeled data				
ACC / STD (%)	100.0 / 00.00	94.16 / 08.52	100.0 / 00.00	98.91 / 02.62
Time Cost (s)	267.21	105.48	267.21	139.31
Model Parameters: 1.52 M				

Table 15: Model performance and computational complexity of SCMM when fine-tuning with limited labeled data (1%, 5%, 10%, and 20%) on SEED and SEED-IV.

Decoding Human Emotions: Analyzing Multi-Channel EEG Data using LSTM Networks

Shyam K Sateesh¹, Sparsh BK¹, and Uma D.¹

¹PES University, Bengaluru, India,
shyamksateesh@gmail.com

Abstract. Emotion recognition from electroencephalogram (EEG) signals is a thriving field, particularly in neuroscience and Human-Computer Interaction (HCI). This study aims to understand and improve the predictive accuracy of emotional state classification through metrics such as valence, arousal, dominance, and likeness by applying a Long Short-Term Memory (LSTM) network to analyze EEG signals. Using a popular dataset of multi-channel EEG recordings known as DEAP, we look towards leveraging LSTM networks' properties to handle temporal dependencies within EEG signal data. This allows for a more comprehensive understanding and classification of emotional parameter states. We obtain accuracies of 89.89%, 90.33%, 90.70%, and 90.54% for arousal, valence, dominance, and likeness, respectively, demonstrating significant improvements in emotion recognition model capabilities. This paper elucidates the methodology and architectural specifics of our LSTM model and provides a benchmark analysis with existing papers.

Keywords: EEG, Emotion Recognition, LSTM, Neural Networks, Deep Learning, Valence, Arousal, Dominance, Likeness, HCI

1 Introduction

EEG is defined as the electrical activity of an alternating type recorded from the scalp surface after being picked up by metal electrodes and conductive media [19]. The unique ability of EEG signals to provide a very descriptive temporal view of brain activity makes it an indispensable tool for understanding complex human emotional states. This capability is especially critical in contexts where the traditional means of emotion assessment are impractical or unfeasible.

In recent years, there has been a necessity for understanding and quantifying emotional responses, which has led to advancements in academic research. This has opened new doors for consumer research, mental health, and assistive technologies. The prospect of its ability to assist individuals who would otherwise not be able to express emotions through traditional ways, such as facial expressions, body language, and speech, makes this one of the exciting fields for EEG-based recognition of emotions. These individuals would include, but not be limited to, people with communication disabilities, for example, aphasia; other

conditions encompass Autism Spectrum Disorder (ASD) [7], among others; and those who have severe physical disabilities from traumatic brain injuries or other progressive diseases, such as Amyotrophic Lateral Sclerosis (ALS) [3].

Statistically, it has been estimated that in the United States alone, approximately 6.6 million people have been diagnosed with some communication disorder [14]. From such figures in the global context, it can be estimated that up to 1% of the world population has some form of autism spectrum disorder. These people's emotional expression and interpretation remain very conventional, usually exhibiting a failing nature. Such failures, thus, stimulate the need to develop standalone technologies that will independently interpret emotional states from physiological data.

EEG-based technologies offer a non-invasive, more direct window into the neural underpinnings of emotion. Since it measures electrical activity, the EEG provides a dynamic mapping of activity in the brain, potentially associated with states of emotion without the need for verbal reports or precise physical gestures. This approach is particularly suitable for those whose neurological conditions impair their effective communication.

The rise of Long Short-Term Memory (LSTM), a variation of Recurrent Neural Networks (RNNs), has revolutionized this field with its ability to analyze and classify EEG data at unprecedented success rates. LSTMs, in particular, are very strong at modelling the time-dependent features that underlie EEG data. They can capture such underlying patterns temporally, which indicate different emotional states. This advanced machine learning method elevates the prediction performance for emotion classification systems from EEG. It opens an avenue to building real-time responsive systems that can adapt to the emotional feedback of users in different applications.

Furthermore, emotion recognition by EEG is possible in healthcare and societal applications. In healthcare, technology could offer better patient care, for instance, since it can interpret pain, discomfort, or emotional distress that patients might be unable to express. One case for this was gauging emotional states in palliative care cancer patients [17]. In special education, EEG in non-verbal students could help teachers and caregivers explore the thoughts and emotional states of the students. This could enable tailor-made educational approaches that are more in sync with the mindset of the students [18].

This study aims to develop more accurate and specific tools for cognitive emotion recognition, particularly for detecting and interpreting emotional states in persons unable to express themselves by traditional means.

2 Related Work

The DEAP dataset, detailed by Koelstra et al. [10], has been foundational in the field, providing a rich data source for subsequent research. Also, significant correlations were found between the participant ratings and EEG frequencies. The single-trial classification was performed for arousal, valence, and liking scales using features extracted from the EEG, peripheral, and MCA modalities. The results were shown to be significantly better than random classification.

Alhagry et al. [2] proposed LSTM networks for emotion recognition from raw EEG signals. Their study demonstrates that the LSTM model achieves high average accuracies across three emotional dimensions and outperforms traditional emotion recognition techniques, marking a significant advancement in the field.

Nie et al. [16] explore the relationship between EEG signals and emotional responses while watching movies, focusing on classifying emotions into positive and negative categories. Their application of a Support Vector Machine (SVM) on processed EEG features resulted in an impressive average testing accuracy of 87.53%, underscoring the potential of EEG-based methods in practical multimedia applications.

Li et al. [11] provide a comprehensive overview of EEG-based emotion recognition, exploring the integration of psychological theories with physiological measurements. They review various machine learning techniques, from conventional models to advanced computational methods, highlighting key advancements and challenges in the field.

Zheng et al. [21] develop an innovative approach by integrating deep belief networks with hidden Markov models for EEG-based emotion classification. Their findings indicate that this combined DBN-HMM model achieves higher accuracy than traditional classifiers, highlighting its effectiveness in leveraging spatial and temporal EEG data dimensions.

Bhagwat et al. [6] proposed a novel approach for classifying four primary emotions: happy, angry, crying, and sad, which can be visualized as four quadrants. They used Wavelet Transforms (WT) to extract features from raw EEG signals and employed a Hidden Markov Model (HMM) to classify emotions.

Lin et al. [13] utilize EEG data and machine learning to enhance emotional state predictions during music listening. Using an SVM, their approach achieves an average classification accuracy of 82.29% for emotions such as joy, anger, sadness, and pleasure.

Naser and Saha [15] applied advanced signal processing techniques to improve feature extraction for emotion classification from EEG signals. Their study utilizes dual-tree complex wavelet packet transform (DT-CWPT) and statistical

methods like QR factorization and singular value decomposition (SVD) to select discriminative features effectively. The enhanced feature set is then classified using an SVM, demonstrating notable improvements in classification accuracy.

Li et al. [12] found that while it is feasible to work with single-channel EEG data, it is much more effective to combine multiple channels of EEG features into a single feature vector. They also found that the beta and gamma frequency bands are more related to emotional processing than the other bands.

The exploration of adaptive emotion detection using EEG and the Valence-Arousal-Dominance model by Gannouni et al. [8] advances the field by adapting computational models to individual brain activity variations. Their method employs an adaptive selection of electrodes, significantly enhancing emotion detection accuracy. Utilizing machine learning algorithms, the study demonstrates a 5% and 2% increase in accuracy for valence, arousal, and dominance dimensions, respectively, compared to fixed-electrode approaches.

Alvarez-Jiménez et al. [22] enhance EEG-based emotion recognition by integrating diverse feature sets from multiple domains. Their use of various classifiers, including Artificial Neural Networks, achieves a high accuracy of 96%, demonstrating the effectiveness of hybrid features in improving model robustness.

Atkinson-Abutridy et al. [5] proposed a feature-based emotion recognition model combining statistical-based feature selection methods with SVM classifiers, focusing on Valence/Arousal dimensions for classification. This combined approach outperformed other recognition methods.

Yoon and Chung [20] detailed a probabilistic classifier based on Bayes' theorem and a supervised learning approach using a perceptron convergence algorithm, offering a methodologically distinct perspective on emotion classification from EEG signals.

3 Dataset

The Database for Emotion Analysis using Physiological Signals (DEAP) [10] is at the core of our study, and it presents a rich source of EEG and peripheral physiological signals for analyzing emotions. The dataset was built to boost and proliferate the development of systems that would be capable of recognizing human emotions from physiological responses, with particular emphasis on the paradigms of human-computer interaction.

3.1 Dataset Description

The DEAP dataset consists of EEG data recordings from 32 participants between 19 and 37 years old, with a mean age of 26.9 years. Each participant was

presented with 40 one-minute music video clips to elicit emotional responses. Participants had to rate their experience after each stimulus on a 1 to 9 integer scale for arousal, valence, dominance, and liking. We will use these subjective ratings as labels to train our models.

3.2 Data Acquisition

EEG and peripheral physiological signals were acquired simultaneously, viewing each music video clip by all the participants. In the course of the experiment, the recording of the EEG data was carried out at 512 Hz through the 32-channel systems, which was eventually reduced to 128 Hz during analysis. Concurrently with EEG, other physiological signals such as galvanic skin response and heart rate were also recorded to deliver complete states regarding the participant's physiological states during each trial.

3.3 Data Structure

For each participant, it is composed of two main arrays: the EEG signals and an array of labels for each trial. The EEG data array has a dimension of 40x40x8064 for 40 trials, 40 channels, and 8064 data points per channel per trial. Corresponding to four emotional dimensions assessed per video clip, the array structure of labels is 40x4. (Shown in Table 1)

Table 1: DEAP: Structure of each participant array

Array Name	Array Shape	Array Contents
data	40 x 40 x 8064	video/trial x channels x data
labels	40 x 4	video/trial x label (VADL)

4 Valence, Arousal, and Dominance Model

The Valence-Arousal-Dominance (VAD) model presents a sensitive framework for recognizing human emotions and classifies them into three significant aspects: valence, arousal, and dominance. Valence measures the 'how good or bad' of the mood, arousal measures the activation level, and dominance measures how much control one might feel they have over their emotional state. Researchers have adopted this model in examining features of EEG that indicate the functioning of different areas in the brain toward emotional stimuli. Research has shown that the positive effect increases alpha band activity in the frontal regions. In contrast, the negative one tends to decrease it, and high arousal corresponds to

beta activity.

The usage of recent advanced classification techniques, even with segmenting EEG data into minor epochs, has increased the accuracy of emotional assessments. Such improvements enhance the accuracy by around 5% for valence and arousal and 2% for dominance, respectively, hence depicting the effectiveness of the VAD model in the subtle multi-dimensionality characterization of human emotions [8]. (VAD Model Shown in Fig. 1 [8])

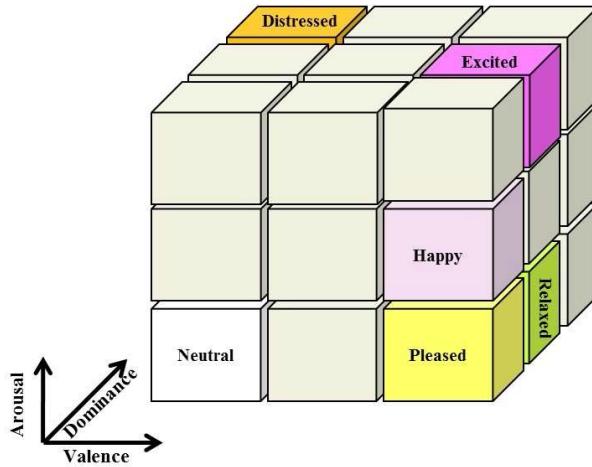


Fig. 1: Valence-Arousal-Dominance Model Depiction as A 3-D Graph [8]

5 Rationale Behind Usage of Deep Learning

Emotion recognition from EEG data is a difficult task due to the complexity and variability of the signals. Although traditional statistical methods effectively analyze structured and more straightforward datasets, they frequently fail to capture and interpret the dynamic and non-linear interactions typical of EEG data. Deep learning, a branch of machine learning, has become an invaluable tool for managing these complexities due to its ability to discern high-level, abstract features from vast amounts of data.

5.1 Deep Learning vs. Traditional Statistical Methods

Deep learning models that handle unstructured data, like images, speech, or biological signals, perform this function due to their use of neural networks. Traditional statistical approaches to data analysis require manual choice of features, and at most, they can only model the linear effects. This is essential in EEG

data, where emotional states are not explicitly encoded but latent constructs reflected in slight signal variations. Deep learning models enable learning such patterns directly from raw data, optimizing feature extraction, selection, and classification tasks in a joint form. This shows that robust and accurate analysis is developed in high dimensionality and noise levels that are usually related to EEG recordings.

5.2 Overview of LSTMs

LSTMs are one of the unique variants of Recurrent Neural Networks (RNN). It was first introduced by Hochreiter et al. [9] to eliminate the problem of long-term dependencies seen in conventional RNNs. Traditional RNNs are known to also suffer from gradient-related issues. This problem, in turn, makes it very hard for them to be trained on sequential data where long-term contextual information is essential. LSTMs solve this problem due to the exceptional structure of their gates, which allows them to regulate the flow of information in a way that enables them to remember or forget information for long periods.

5.3 Bidirectional LSTMs

The capabilities of standard LSTMs are further advanced through the usage of bidirectional LSTMs, allowing more context to be available from the subsequent points in the data sequence. So, bidirectional LSTMs can capture the context information from past and future states by processing data in the forward and reverse directions. This is very useful in emotion recognition from the EEG signals when the emotional state reflected in a data segment may depend not only on the earlier but also on the latter events.

5.4 LSTM for Our Work

In this project, we choose LSTM networks due to their prowess in sequence prediction problems, thus capable of adequately modelling the temporal dynamics characteristic of EEG data. Applying LSTMs will help reach the deepest emotional timelines that fall within the EEG signals, making them more helpful in predicting emotional states with better accuracy. The bidirectional approach of the capability enforces the complete context from all the data points, which increases the recognition accuracy for complex emotional states. This makes LSTMs very apt for the development of a robust system for emotion recognition from EEG-based data.

6 Proposed Method

Our study uses an LSTM model to classify emotional states from EEG data and focuses on feature extraction, data preparation, and architectural considerations to achieve high accuracy percentages.

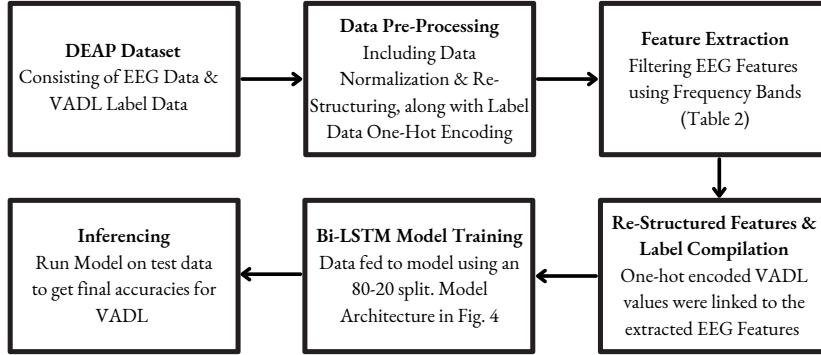


Fig. 2: Flowchart of Proposed Scheme

6.1 Pre-Processing Methods Used

The feature extraction process was tailored to capture significant information from the EEG signals. We utilized specific EEG channels and frequency bands relevant to emotional processing. The chosen channels included a subset correlating with emotional states, such as frontal and temporal regions. Frequency bands were segmented into five distinct ranges: theta (4-8 Hz), alpha (8-12 Hz), low beta (12-16 Hz), high beta (16-30 Hz), and gamma (30-45 Hz), which are traditionally associated with different aspects of cognitive processing and emotional regulation (Refer to Table 2 and Fig. 3 [4]). Each of these bands aids in extracting vital information from input EEG data, which has been proven to support sentiment analysis [4]. The Fast Fourier Transform (FFT) process was applied to a select 14 channels of the recorded 32 channels, chosen to fit Emotiv Epoc, with a window size of 256 points, corresponding to 2 seconds of data, with an overlap of 0.125 seconds to ensure comprehensive temporal analysis.

Table 2: EEG Feature Bands Used in the Study

Brainwave Type	Frequency Range (Hz)	Mental States & Conditions Seen
Theta	4 - 8	Intuitive, creative, recall, fantasy, imaginary, dream
Alpha	8 - 12	Relaxed but not drowsy, calm, conscious
Low-Beta	12 - 16	Relaxed yet focused, integrated
High-Beta	16 - 30	Alertness, agitation
Gamma	30 - 45	Cognition, information processing

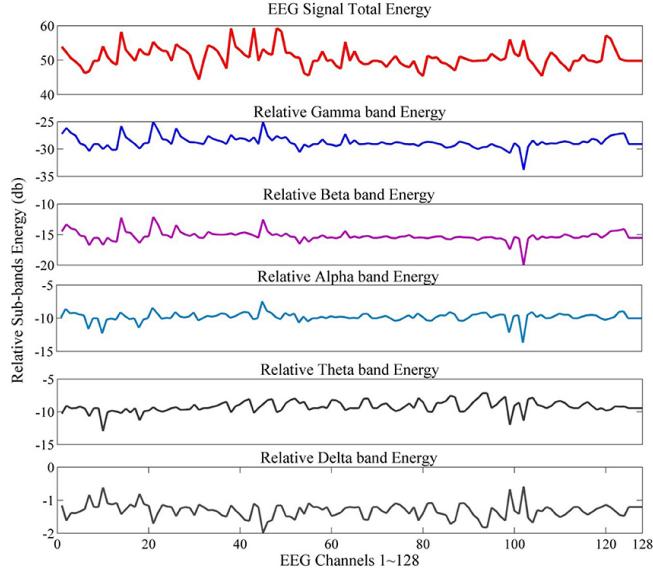


Fig. 3: EEG signal energy and relative sub-band energy [4]

The dataset was first split into training and test splits using an 80-20 ratio. That is, 80% of the data was used to train the LSTM model, and the remaining 20% was utilized to test the model's performance. This split ratio enabled the practical training of the model as well as a reliable evaluation to establish how it generalizes to unseen data.

Data Normalization was necessary to normalize the input features to reduce discrepancies in signal amplitudes caused by data variations across individuals. All feature vectors were normalized to zero mean and unit variance, a standard approach in processing EEG signals to overcome inter-subject differences.

The next step was converting each valence, arousal, dominance, and likeness label (initially scaled from 1-9) into one-hot encodings to create nine classes before sending the label data into the LSTM Networks. This was implemented using the `keras.utils.to_categorical()` function.

6.2 LSTM Architecture

The LSTM network architecture employed in our study is designed to handle EEG data sequentially and temporally effectively. We will use one LSTM model for each emotional parameter in observation. The model initiates with a Bidirectional LSTM layer consisting of 128 units, enhancing the model's ability to capture dependencies in both forward and backward directions of the input sequence. This layer is followed by a dropout of 0.6 to reduce overfitting by randomly ignoring a fraction of the neurons during training.

Subsequent layers include multiple LSTM layers with varying numbers of neurons to extract and refine features from the data incrementally. Specifically, the model includes an LSTM layer with 256 units and another two LSTM layers, each with 64 units, all incorporating a dropout of 0.6 after each LSTM layer to prevent overfitting further. The final LSTM layer consists of 32 units, followed by a dropout of 0.4, aiming to consolidate the features extracted by previous layers into a more manageable form.

The output from the LSTM layers is then passed through two dense layers. The first dense layer has 16 units with a ReLU activation function intended to introduce non-linearity into the model, facilitating the network's ability to learn complex patterns. The final output layer consists of some units equal to the classes of emotions being classified, with a softmax activation function to output the probability distribution over the classes.

This architecture is compiled with the Adam optimizer and categorical cross-entropy as the loss function, suitable for multi-class classification problems. The detailed structure and parameterization of the model are crucial for its ability to discern nuanced emotional states from EEG data, as visualized in the accompanying architectural diagram in our study. A representation of the model is shown in Fig. 4.

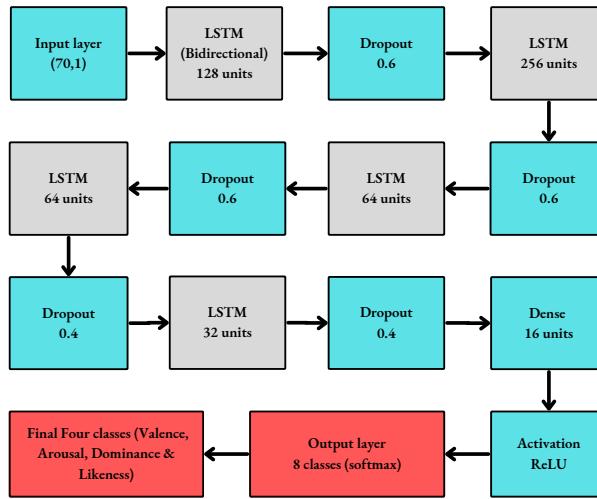


Fig. 4: Our LSTM Architecture

7 Results

Our LSTM-based model demonstrated outstanding performance in emotion recognition from EEG data, achieving individual class accuracies of 90.33% for valence, 89.89% for arousal, 90.70% for dominance, and 90.54% for likeness, with an overall accuracy of 90.36%. These results underline the model's efficacy in capturing complex emotional states through advanced feature extraction and a robust LSTM architecture. This performance showcases the model's capabilities and sets a foundation for future advancements in EEG-based emotion recognition. A comparison of accuracies is attached in Table 3, showing our method of using an LSTM network to be highly accurate and effective in classifying emotional parameters correctly compared to related papers.

Table 3: Average Accuracies and Nature of Features Extracted

Paper	Arousal	Valence	Liking	Features Extracted	Frequency Bands Ranges (Hz)
Koelstra et al. [10]	62.00	56.70	55.40	Frequency Based	(4-7), (8-13), (14-29), (30-47)
Atkinson and Campos [5]	73.06	73.41	—	Statistical Based	—
Yoon and Chung [20]	70.10	70.90	—	Frequency Based	(4-8), (8-13), (13-30), (36-44)
Naser and Saha [15]	66.20	64.30	70.20	Time-Frequency Based	—
Alhagry et al. [2]	85.65	85.45	87.99	Frequency Based	(4-8), (8-10), (8-12), (12-30), (30+)
Li et al. [11]	83.78	80.72	—	Frequency Based	(4-8), (8-13), (13-30), (30-45)
Acharya D et al. [1]	—	—	88.60	Frequency Based	(4-8), (8-12), (12-16), (16-25), (25-45)
Proposed Method	89.89	90.33	90.54	Frequency Based	(4-8), (8-12), (12-16), (16-30), (30-45)

8 Conclusion

This study successfully showcases the efficacy of LSTM networks in accurately classifying emotional states from EEG data, achieving high performance across various emotional dimensions. The customized LSTM architecture, incorporating bidirectional layers and strategic dropout stages, adeptly handles the complexities of EEG signals. Our LSTM architecture paired with uniform frequency band ranges taken for EEG feature extraction has proven to provide improved results from previous LSTM-based EEG studies. Such capabilities pave the way for advancements in cognitive neuroscience and human-computer interaction, promising enhancements in responsive systems that adapt to user emotions in real time. Future work can further build upon this model with more robust neural networks, including time-frequency and location domain features, along with the possible usage of more than 14 EEG channels for better efficiency of emotion recognition. Upcoming research will benefit from exploring hybrid models that integrate additional physiological signals, further refining the precision and application of EEG-based emotion recognition in creating empathetic user interfaces.

References

1. Divya Acharya, Riddhi Jain, Siba Smarak Panigrahi, Rahul Sahni, Siddhi Jain, Sanika Prashant Deshmukh, and Arpit Bhardwaj. Multi-class emotion classification using eeg signals. In Deepak Garg, Kit Wong, Jagannathan Sarangapani, and Suneet Kumar Gupta, editors, *Advanced Computing*, pages 474–491, Singapore, 2021. Springer Singapore.
2. Salma Alhagry, Aly Aly Fahmy, and Reda A. El-Khoribi. Emotion recognition based on eeg using lstm recurrent neural network. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(10), 2017.
3. Manouchehr Shamseini Ghiyasvand Alireza Pirasteh and Majid Pouladian. Eeg-based brain-computer interface methods with the aim of rehabilitating advanced stage als patients. *Disability and Rehabilitation: Assistive Technology*, 0(0):1–11, 2024. PMID: 38400897.
4. Hafeez Ullah Amin, Wajid Mumtaz, Ahmad Rauf Subhani, Mohamad Naufal Mohamad Saad, and Aamir Saeed Malik. Classification of eeg signals based on pattern recognition approach. *Frontiers in Computational Neuroscience*, 11, 2017.
5. John Atkinson and Daniel Campos. Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers. *Expert Systems with Applications*, 47:35–41, 2016.
6. Anuja R. Bhagwat and A. N. Paithane. Human disposition detection using eeg signals. In *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, pages 366–370, 2016.
7. Nash N. Boutros, Renee Lajiness-O'Neill, Andrew Zillgitt, Anette E. Richard, and Susan M. Bowyer. Eeg changes associated with autistic spectrum disorders. *Neuropsychiatric Electrophysiology*, 1(1):3, 2015.
8. Sofien Gannouni, Arwa Aledaily, Kais Belwafi, and Hatim Aboalsamh. Adaptive emotion detection using the valence-arousal-dominance model and eeg brain rhythmic activity changes in relevant brain lobes. *IEEE Access*, 8:67444–67455, 2020.
9. Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.
10. Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.
11. Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. Eeg based emotion recognition: A tutorial and review. *ACM Comput. Surv.*, 55(4), nov 2022.
12. Li, Xian, Yan, Jian-Zhuo, and Chen, Jian-Hui. Channel division based multiple classifiers fusion for emotion recognition using eeg signals. *ITM Web Conf.*, 11:07006, 2017.
13. Yuan-Pin Lin, Chi-Hong Wang, Tien-Lin Wu, Shyh-Kang Jeng, and Jyh-Horng Chen. Eeg-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 489–492, 2009.
14. Megan A. Morris, Sarah K. Meier, Joan M. Griffin, Megan E. Branda, and Sean M. Phelan. Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey. *Disability and Health Journal*, 9(1):140–144, 2016.

15. Daimi Syed Naser and Goutam Saha. Recognition of emotions induced by music videos using dt-cwpt. In *2013 Indian Conference on Medical Informatics and Telemedicine (ICMIT)*, pages 53–57, 2013.
16. Dan Nie, Xiao-Wei Wang, Li-Chen Shi, and Bao-Liang Lu. Eeg-based emotion recognition during watching movies. In *2011 5th International IEEE/EMBS Conference on Neural Engineering*, pages 667–670, 2011.
17. Rafael Ramirez, Josep Planas, Nuria Escude, Jordi Mercade, and Cristina Farriols. Eeg-based analysis of the emotional effect of music therapy on palliative care cancer patients. *Frontiers in Psychology*, 9, 2018.
18. Nazmi Sofian Suhaimi, James Mountstephens, and Jason Teo. Eeg-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational Intelligence and Neuroscience*, 2020(1):8875426, 2020.
19. Michal Teplan. Fundamental of eeg measurement. *MEASUREMENT SCIENCE REVIEW*, 2, 01 2002.
20. Hyun Joong Yoon and Seong Youb Chung. Eeg-based emotion estimation using bayesian weighted-log-posterior function and perceptron convergence algorithm. *Computers in Biology and Medicine*, 43(12):2230–2237, 2013.
21. Wei-Long Zheng, Jia-Yi Zhu, Yong Peng, and Bao-Liang Lu. Eeg-based emotion classification using deep belief networks. volume 2014, 07 2014.
22. Mayra Álvarez Jiménez, Tania Calle-Jimenez, and Myriam Alvarez. A comprehensive evaluation of features and simple machine learning algorithms for electroencephalographic-based emotion recognition. *Applied Sciences*, 14:2228, 03 2024.

Emotion-Agent: Unsupervised Deep Reinforcement Learning with Distribution-Prototype Reward for Continuous Emotional EEG Analysis

Zhihao Zhou^{1,2}, Qile Liu^{1,2}, Jiyuan Wang^{1,2}, Zhen Liang^{1,2,*}

¹School of Biomedical Engineering, Shenzhen University, Shenzhen, 518060, Guangdong, China

²Guangdong Provincial Key Laboratory of Biomedical Measurements and Ultrasound Imaging, Shenzhen, China
{2310247057, liuqile2022, 2310247016}@email.szu.edu.cn, janezliang@szu.edu.cn

Abstract

Continuous electroencephalography (EEG) signals are widely used in affective brain-computer interface (aBCI) applications. However, not all continuously collected EEG signals are relevant or meaningful to the task at hand (e.g., wondering thoughts). On the other hand, manually labeling the relevant parts is nearly impossible due to varying engagement patterns across different tasks and individuals. Therefore, effectively and efficiently identifying the important parts from continuous EEG recordings is crucial for downstream BCI tasks, as it directly impacts the accuracy and reliability of the results. In this paper, we propose a novel unsupervised deep reinforcement learning framework, called Emotion-Agent, to automatically identify relevant and informative emotional moments from continuous EEG signals. Specifically, Emotion-Agent involves unsupervised deep reinforcement learning combined with a heuristic algorithm. We first use the heuristic algorithm to perform an initial global search and form prototype representations of the EEG signals, which facilitates the efficient exploration of the signal space and identify potential regions of interest. Then, we design distribution-prototype reward functions to estimate the interactions between samples and prototypes, ensuring that the identified parts are both relevant and representative of the underlying emotional states. Emotion-Agent is trained using Proximal Policy Optimization (PPO) to achieve stable and efficient convergence. Our experiments compare the performance with and without Emotion-Agent. The results demonstrate that selecting relevant and informative emotional parts before inputting them into downstream tasks enhances the accuracy and reliability of aBCI applications.

Introduction

Human emotion is a continuous dynamic process, characterized by complex interactions between both internal and external components of the human body (Cowen and Keltner 2017; Horikawa et al. 2020). How to identify task-related emotional segments from continuous EEG signals presents a significant challenge. Electroencephalography (EEG) provides a direct, objective, and scientifically grounded method for assessing emotional states, making it a valuable tool in emotion recognition research (Song et al. 2018). In recent years, the potential of EEG-based emotion recognition

has garnered increasing attention from researchers across diverse disciplines (Li, Wang, and Lu 2021; Gong et al. 2023; Liu et al. 2024).

One significant limitation of existing research is the reliance on a static labeling approach, where a single, fixed label is assigned to an entire EEG segment. This method fails to capture the dynamic nature of human emotions during EEG-evoked experiments, as emotional states are inherently fluid, constantly shifting in response to both internal cognitive processes and external stimuli (Huang et al. 2014; Liu et al. 2017). Moreover, continuous EEG recordings often include states that are irrelevant to the specific task being studied. These irrelevant states can introduce noise and confounding factors, undermining the accuracy and reliability of emotion recognition models. Current methods face challenges in isolating and identifying the task-related moments within the EEG data that are most relevant to the study. When task-irrelevant EEG segments are included in the training data, they introduce extraneous information that can degrade the model's performance. As a result, the model may mistakenly associate these irrelevant patterns with emotional states, leading to reduced accuracy in emotion recognition by diverting attention from the true task-related emotional dynamics (Li et al. 2019; Zheng and Lu 2015; Zheng 2016). On the other hand, requiring real-time annotation of task-related segments during an experiment is impractical. This is especially true when considering that wandering thoughts or irrelevant mental states are often indistinguishable even to the subject themselves. Thus, developing an artificial intelligence (AI) empowered method that can dynamically adapt to the fluid nature of human emotions and accurately isolate task-relevant EEG segments is essential for improving the precision and effectiveness of emotion recognition models.

Deep reinforcement learning, with its adaptability and flexibility in uncertain environments, offers a promising solution to this challenge (Vinyals et al. 2019; Kalashnikov et al. 2018). By leveraging a reward-based mechanism, it reduces the dependence on labels and enables unsupervised autonomous exploration of task-relevant information. For example, Zhou (Zhou, Qiao, and Xiang 2018) proposed a Diversity-Representativeness Reward to guide Agent in generating more diverse and representative video summaries. Similarly, AC-SUM-GAN (Apostolidis et al. 2020) used an

*Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Actor-Critic framework to exploit the reconstruction error of the discriminator as a reward function, with the Critic guiding the Actor through gradient feedback to learn strategies for extracting key video segments. In the field of EEG emotion computing, TAS-Net (Zhang et al. 2023) proposed the use of deep reinforcement learning to detect the most informative key emotional segments from EEG signals in an unsupervised manner. However, these methods often fail to incorporate information about the overall distribution of EEG signal features and struggle to capture the long-term similarities in human emotions. This can lead to gaps in understanding the continuity and subtle shifts in emotional states, potentially affecting the accuracy and effectiveness of emotion recognition models.

To address the limitations of existing research, we formulate the extraction of key EEG segments as a sequential decision-making process and introduce a novel Emotion-Agent designed to automatically identify relevant and informative emotional moments from continuous EEG signals. Emotion-Agent integrates reinforcement learning with heuristic search algorithms to enhance the RL agent's exploration process during training. By utilizing the efficient search capabilities of heuristic algorithms, the agent can minimize exploration of low-value trajectories, making the process more targeted and purposeful. Consequently, the model achieves more efficient convergence, even when accounting for the inherent costs of exploration. The proposed Emotion-Agent is capable of effectively capturing the most significant emotional segments without the need for predefined labels. The main contributions of this paper are summarized as follows.

- We propose a novel Emotion-Agent, which integrates deep reinforcement learning with heuristic algorithms to optimize the extraction of key emotional segments.
- The reward function, named Distribution-Prototype, is designed with a focus on distribution, considering both local and global sample distributions during the reward learning process.
- Extensive experimental results demonstrate that selecting relevant and informative emotional segments enhances the accuracy and reliability of emotion analysis.

Related Work

Reinforcement Learning

Reinforcement Learning (RL) is a powerful machine learning paradigm where an intelligent agent learns an optimal decision policy by interacting with its environment (Zoph and Le 2017). Unlike other machine learning methods, RL emphasizes learning through trial and error, with the agent taking actions to maximize cumulative rewards over time. This approach has gained significant traction across various domains due to its ability to handle complex, dynamic environments where the agent's decisions continuously adapt based on new information (He et al. 2016; Yarats, Kostrikov, and Fergus 2021).

Reinforcement Learning with Heuristics

Heuristic-Guided Reinforcement Learning (HuRL) was introduced (Cheng, Kolobov, and Swaminathan 2021), aiming to accelerate traditional RL algorithms by incorporating heuristics derived from domain knowledge or offline data. These heuristics guide the RL agent, enabling more informed decisions and speeding up the learning process. HuRL is particularly valuable in environments where the state space is vast, making unguided exploration computationally expensive and time-consuming. Another significant advancement is the introduction of large-state reinforcement learning for hyper-heuristics (Kletzander and Musliu 2023). This approach leverages solution change trajectories from an extensive feature set, integrating them into the RL framework. By incorporating local search principles and introducing a probability distribution within the ϵ -greedy strategy, this method increases the likelihood of sampling high-quality sequences of low-level heuristics. It significantly enhances the efficiency of RL in solving complex optimization problems with exceptionally large state spaces. A novel solution for continuous trajectory generation in urban road networks was also proposed (Jiang et al. 2023), combining a two-stage Generative Adversarial Network (GAN) with A* heuristic search algorithms. This design features discriminators for sequential reward and movement yaw reward, guiding the agent in generating more accurate and efficient trajectories. Building on the foundations of RL, personalized reinforcement learning was introduced (Ivanov and Ben-Porat 2024). Inspired by the classical K-means clustering principle, this approach incorporates the concept of a budget of policies within robust Markov Decision Processes (r-MDPs). The framework enables the RL agent to interact with users through representative policies, efficiently adapting to individual user preferences. An earlier application of RL in the field of education is demonstrated with AgentX (Martin and Arroyo 2004). This intelligent agent was developed to enhance the effectiveness of Intelligent Tutoring Systems (ITS). By clustering personalized group information about students, the RL-based AgentX tailors the learning experience for each group.

The advancements in reinforcement learning across various domains: from accelerating traditional RL algorithms with heuristics to personalizing user interactions and improving intelligent tutoring systems—demonstrate the versatility and potential of this machine learning paradigm. The continuous evolution of RL, as seen in large-state optimization and urban trajectory generation, underscores its capacity to tackle increasingly complex challenges. As RL continues to integrate with other AI techniques, such as GANs and heuristic search algorithms, it is positioned to drive significant innovations across a wide array of fields, shaping the future of intelligent systems and autonomous decision-making (Warnell et al. 2018; Ren et al. 2023; Vecerik et al. 2017).

Methodology

Markov Decision Process

We model the detection of the most emotionally rele-

vant segments from sequential EEG signals as a sequential decision-making process, formulated as a Markov Decision Process (MDP). An MDP is defined as a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, where \mathcal{S} represents the state space, and \mathcal{A} denotes the action space, with $\mathcal{A} = \{0, 1\}$ corresponding to the possible actions the agent can take. The transition probability function $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ describes the likelihood of transitioning from one state to another given a specific action. The reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathcal{R}$ assigns a numerical reward based on the state-action-state transition, providing feedback on the agent's decisions.

At each timestep t , the RL agent, upon executing an action a_t in state s_t , transitions to a new state s_{t+1} and receives a corresponding reward r_t . We define \mathcal{M} tuple as transitions. In an episode when interacting with the environment, we collect multiple trajectories consisting of multiple transitions. Through repeated interactions with the environment, the RL algorithm aims to learn an optimal policy π that maximizes the cumulative reward over time. This optimal policy enables the agent to make decisions that consistently lead to the identification of key emotional segments in the EEG signals.

Prototype Learning

In the process of extracting key segments from task-related EEG signals, it is crucial to consider not only the intrinsic data distribution characteristics of each segment but also the broader context provided by the global emotion space, which encapsulates the distribution of various emotion categories. To achieve a more effective representation of EEG emotion distribution, we introduce the concept of prototype learning (Zhou et al. 2023). This approach allows us to model each emotion category as a prototype, thereby capturing the globally distributed emotional information with better representation. Prototype learning enables us to integrate global emotion information into the reward structure, ensuring that the reinforcement learning process is informed by a comprehensive understanding of the emotional landscape represented in the EEG data.

Specifically, we employ the K-Means clustering algorithm as a heuristic method. This step enables the model to obtain a global perspective on the distribution of emotional information across all subjects' EEG data. By inputting the differential entropy (DE) features of EEG signals from all subjects into the K-Means algorithm, we derive the set of emotion prototypes $\{C^i\}_{i=1}^N$, where each C^i represents a cluster center corresponding to a emotion category, N represents the number categories. By clustering the data, we identify the optimal emotion prototypes that serve as representative points in the emotion space. These prototypes are then used to inform the design of the reward function, ensuring that it reflects the global distribution of emotions captured in the EEG signals. The prototype feature vector for a given emotion category c can be calculated by averaging all the sample features that belong to this category. Mathematically, the prototype feature vector μ_c is given by:

$$\mu_c = \frac{1}{|C^i|} \sum_{x_i \in C^i} f(x_i), \quad (1)$$

where $C^i = \{(x_i, y_i = c)\}_{i=1}^N$ represents the set of samples belonging to the emotion category c , and $|C^i|$ is the number of samples in this category. The centroid μ_c serves as the average feature vector for the emotion category c .

The K-Means algorithm iteratively reclassifies data points and updates cluster centers to minimize the sum of squared errors within the clusters. The objective function of the algorithm is defined as:

$$\arg \min_C \sum_{i=1}^N \sum_{x \in C^i} \|x - \mu_c\|_2. \quad (2)$$

To quantify the variance within each cluster, we use the mean of the sum of squared intra-cluster errors, which reflects the distribution of EEG features within the cluster. The variance of the intra-cluster distribution is indicative of the individual variability of each emotion. The intra-cluster variance σ_i^2 is calculated using the following formula:

$$\sigma_i^2 = \frac{1}{|C^i|} \sum_{x_i \in C^i} \|x_i - \mu_c\|_2, \quad (3)$$

where $|C^i|$ is the number of data points in the cluster, x_i represents an individual data point, and μ_c is the corresponding cluster center.

To better represent the distribution of emotions in the EEG data, the prototype learning process integrates the cluster centers as the mean of the data distribution and uses the mean of the sum of squared errors within clusters to describe the variance. This combination of prototype learning and K-Means clustering provides a robust foundation for the reinforcement learning process, enabling the model to effectively navigate and interpret the complex emotional information present in EEG data.

Distribution-Prototype Reward

We obtain a global distribution of emotional information in an unsupervised manner through prototype learning, and we incorporate this global information into the reward function for our reinforcement learning model. We believe that the clustering centers obtained through heuristic search represent the prototypes of each affective category. These prototypes effectively capture the distribution of sample features across the entire affective category, with other EEG features belonging to the same category clustering around these prototypes.

From a probabilistic perspective, an emotion prototype can be understood as the mean of the emotion sample features, while the variance in this distribution arises from the inherent variability of human emotions and the non-stationarity of EEG signals. We use the mean and variance of each emotion cluster to reflect the individual variability within the distribution of EEG features for each emotion category. The goal of the Actor in our proposed method is to maximize the expected reward over time by selecting key segments that are more closely aligned with the target emotion. To achieve this, we propose two reward functions based on the distributional information derived from prototype learning: center reward and inter-intra reward.

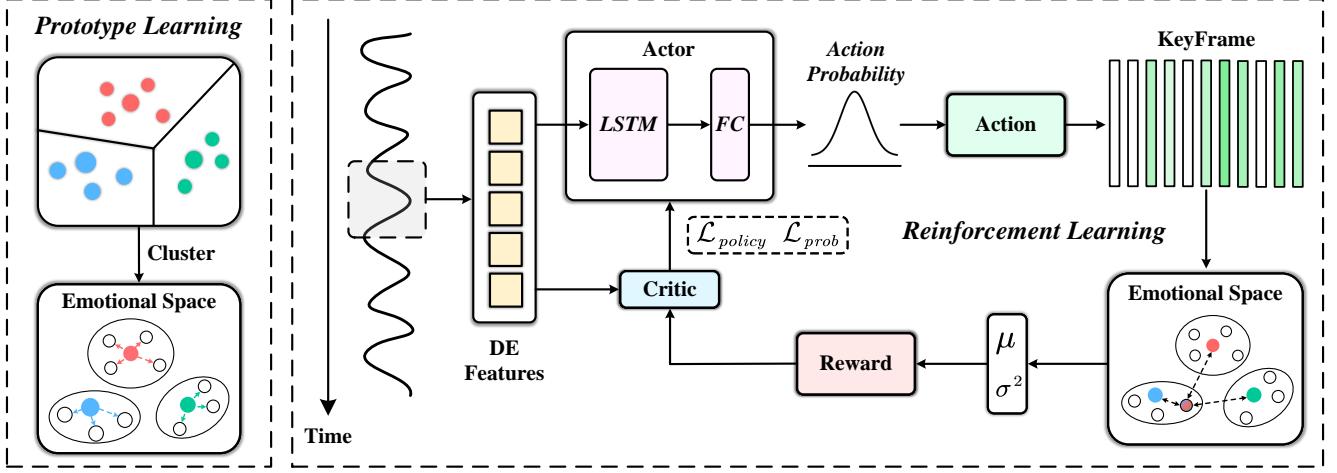


Figure 1: The framework of the proposed Emotion-Agent. The Emotion Agent is divided into two stage in total: (1) Heuristic global search for prototype learning stage, and (2) Reinforcement learning stage based on Distribution-Prototype reward

Center reward This reward function measures the intensity of the emotional information contained in the features of the current EEG sample. We evaluate this intensity by calculating the Euclidean distance between the sample point and the cluster center:

$$Reward_{center} = \frac{1}{1 + dist}, \quad (4)$$

Where $dist$ calculate the distance from the EEG sample feature to the clustering centre of the category it belongs to.

$$dist = \|x_{c^i} - \mu_{c^i}\|_2. \quad (5)$$

We use this distance measure between sample points and cluster centers in the emotion space for two purposes. First, we consider that the proximity of sample feature points to cluster centers reflects the intensity of the emotion they represent. Second, it helps to mitigate the effect of interfering information from outliers in the EEG signal caused by non-stationarity.

Inter-Intra reward This reward function represents the confidence level that the current EEG sample feature belongs to the specific emotion category. We employ inverse variance weighting to calculate the distances between EEG sample features and the cluster centers of other categories. This method provides a weighted mean with the smallest variance (Lin, Deng, and Pan 2021), which we use to estimates the confidence *Inter* that the sample point is far from the centroid of the other category.

$$Reward_{inter-intra} = \exp\left(-\frac{Intra}{Inter}\right), \quad (6)$$

$$Intra = \|x_{c^i} - \mu_{c^i}\|_2, \quad (7)$$

$$Inter = \frac{\sum_{x_i \notin C^i} \frac{\|x_i - x_{c^i}\|_2}{\sigma_i^2}}{\sum_{x_i \notin C^i} \frac{1}{\sigma_i^2}}, \quad (8)$$

where *Intra* estimates the confidence which the sample point belongs to the target category.

Optimization Process

We use PPO (Schulman et al. 2017) to train our model, as it searches for emotionally relevant EEG key segments at the trial level. Throughout this process, the model learns to identify emotionally prototypical policies in a trial-and-error manner within a discrete action space. To stabilize the training process, we employ PPO-Clip, which restricts the ratio between old and new policies, reducing oscillations and accelerating convergence. PPO is a policy-based Actor-Critic method. To improve sample efficiency in the On-Policy training process, importance sampling is introduced, allowing the model to reuse trajectories multiple times:

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)}, \quad (9)$$

where $r_t(\theta)$ represents the probability ratio between the current and previous policies.

To better estimate cumulative returns, we use Generalized Advantage Estimation (GAE), which provides a more accurate advantage function. The specific expression of GAE is as follows:

$$\hat{A}_t^{GAE(\gamma, \lambda)} = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+1}^V = \delta_t^V + \gamma \lambda \hat{A}_{t+1}^{GAE(\gamma, \lambda)}, \quad (10)$$

where γ is the discount factor and λ is the GAE hyperparameter that controls the trade-off between bias and variance. GAE uses a weighted average of multiple value estimates, and to quickly estimate the advantage at each time step, a recursive calculation is performed, estimating time t from time $t+1$. To further stabilize the training process, we apply PPO-Clip, which limits the changes between the old and new policies. The objective function with clipping can be expressed as:

$$\mathcal{L}_{policy}(\theta) = \hat{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right], \quad (11)$$

where ϵ control the update range of the action probability at each iteration by setting upper and lower thresholds on the ratio of the new and old strategies.

Regularization

To prevent the Actor from selecting too many keyframes during an episode, we introduce a regularization term that constrains the action probability of the learned policy function. This can be expressed as:

$$L_{prob} = \left\| \frac{1}{T} \sum_{t=1}^T p_t - \delta \right\|^2, \quad (12)$$

where δ is a scalar representing the desired proportion of key emotional segments selected. We then use the Adam optimizer to update the parameters θ of the policy function, calculated as:

$$\theta = \theta - \varphi \nabla_\theta (-\mathcal{L}_{policy} + \beta \mathcal{L}_{prob}), \quad (13)$$

where φ is the learning rate, and β is a regularization coefficient.

The Critic network is responsible for estimating $V(S_t)$ during the decision process, which serves as a prediction of the actual discounted cumulative reward throughout the process. This estimation guides the Actor network towards converging on the optimal policy. The error function for the Critic is the mean square error (MSE) between the estimated value and the actual discounted cumulative reward, and is expressed as:

$$L(\theta) = \mathbb{E} \left[(V(s_t) - R_t)^2 \right], \quad (14)$$

where R_t is the discounted cumulative reward at time t . The Critic's predicted value is then used to calculate the MSE loss relative to the actual discounted reward.

During the training process, the model is divided into two parts: Actor and Critic. The Actor continually interacts with the emotion space during decision-making, iteratively searching for an optimal strategy based on the reward mechanism provided by environmental feedback. The Critic guides this exploration process by estimating the cumulative reward for each state, thereby assisting the Actor in finding the optimal strategy.

Experiments

Datasets and Implementation Details

Extensive validation experiments are conducted two publicly available datasets, including SEED (Zheng and Lu 2015) and DEAP (Koelstra et al. 2011). We use DE features as inputs for the model. The details of the dataset and preprocessing will be introduced in the appendix.

The network of actor consists of one layer of LSTM where the number of hidden layer nodes is 128 and two fully connected layers where the hidden nodes are from $256 \rightarrow 128, 128 \rightarrow 2$. The network setup of Critic is one layer of LSTM 128 and two fully connected layers where the hidden layer nodes in the fully connected layers are $256 \rightarrow 128, 128 \rightarrow 1$. actor is optimised by the The optimisation is done by Adam's optimiser and the learning rate

Algorithm 1: The pre-training process of Emotion-Agent.

Input: DE feature sequence $\{s_t\}_{t=1}^T$ from trainging set

Output: Parameters θ of the Emotion-Agent.

- 1: Input All Subjects DE feature data into K-Mean, calculate μ_i according to Eq.(2), calculate σ^2 according to Eq.(4);
 - 2: Initial actor parameters θ_0 , initial critic parameters ϕ_0 ;
 - 3: **for** $i \rightarrow 1, 2, \dots, \xi$ **do**
 - 4: Input a DE feature sequence $\{s_t\}_{t=1}^T$ from training set
 - 5: Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by policy π_k of actor;
 - 6: Compute generalized advantage estimates \hat{A}_t , based on the current value function $V_{\phi k}$;
 - 7: Update the policy of actor by maximizing the PPO-Clip objective function Eq.(11);
 - 8: Update the value function of critic by regression on Mean-Squared Error;
 - 9: **end for**
 - 10: Save the parameters θ of the Emotion-Agent.
-

is set to 1e-4 and Critic's learning rate is done by Adam's optimiser and the learning rate is set to 1e-3. In addition the PPO algorithm has the gamma set to 0.98, the lmbda set to 0.95 and the eps set to 0.2 for the algorithms Eq. (9), Eq. (10), Eq. (11). All experiments are conducted using PyTorch 1.13.1 on an NVIDIA GeForce RTX 3090 GPU. More implementation details and parameter analysis are provided in the appendix.

Evaluation Settings and Metrics

We use two experimental protocols to evaluate our approach.

(1) Cross-Subject: **Subject-Independent, Subject-Level LOOCV**. We use subject-leave-one-out cross-validation to test the performance of our proposed model over cross-subjects. (2) Within-Subject: **Subject-Independent, Video-Level LOOCV**.Based on the above experimental scheme, we extract the most emotion-related segments on the proposed model, which are then used for subsequent model method analysis.

We conducted relevant experiments using the proposed model Emotion-Agent on SEED, DEAP datasets, Emotion-Agent extracted the key segments related to emotions in an unsupervised manner through the guidance of reward function, we used the extracted key segments for downstream task modelling, the experiments compared with and without Emotion-Agent's Accuracy, F1-Scores two evaluation metrics.

Experimental Results

We compare the proposed Emotion-Agent with the current state-of-the-art methods. The comparison results for the three-classified emotion recognition (positive, neutral, negative) task on SEED are given in Table 1, where the methodology and the experimental protocol used are clearly stated. Overall, a supervised learning based approach yields better

Methods	Classification Task	P_{acc}
<i>Supervised</i>	<i>Subject-Dependent</i>	<i>Video-Level LOOCV</i>
GSCCA (Zheng 2016)	Three-Class	82.96
DGCNN (Song et al. 2018)	Three-Class	90.40
RGNN (Zhong, Wang, and Miao 2020)	Three-Class	94.24
<i>Supervised with Transfer Learning</i>	<i>Subject-Independent</i>	<i>Subject-Level LOOCV</i>
BiDANN (Li et al. 2018)	Three-Class	83.28
JDA (Li et al. 2019)	Three-Class	88.28
PR-PL (Zhou et al. 2023)	Three-Class	93.06
<i>Supervised without Transfer Learning</i>	<i>Subject-Independent</i>	<i>Subject-Level LOOCV</i>
JDA (Li et al. 2019) (source domain only)	Three-Class	58.23
<i>Unsupervised</i>	<i>Subject-Independent</i>	<i>Subject-Level LOOCV</i>
EEGFuseNet (Liang et al. 2021)	Three-Class	42.04
TAS-Net (Zhang et al. 2023)	Three-Class	52.99
Emotion-Agent (Ours)	Three-Class	62.31

Table 1: Model performance (%) for cross-subject emotion recognition on the SEED dataset.

emotion recognition performance compared to an unsupervised learning based approach due to the use of label information for modelling, but such a trained model introduces label noise, and the model that completes the training actually learns that it is not really relevant to the emotion. In comparison to the unsupervised approach, the classification accuracy P_{acc} with KNN reached 62.31% after our model extracted the key segments, which is an improvement of 9.32% compared to the results of TAS-Net for emotion recognition with the same KNN classifier. The experimental results show that through a well-designed reward function, our proposed method Emotion-Agent is better able to extract more relevant and richer EEG emotion segments on the SEED dataset, and from the experimental results we further argue this result.

On the other hand, we dichotomised both Arousal and Valence for subjects emotional states on the DEAP dataset. Table 2 gives the experimental results of classifying both labels on the DEAP dataset and comparing them with other methods. Our proposed model performs the same on the task of classifying the emotional intensity of Arousal, and P_{acc} is 14.55% higher than TAS-Net, and comparing some supervised learning methods on Subject-Independent, Subject-Level LOOCV is 2.09% higher than the current SOTA ATDD-LSTM. In addition our proposed method outperforms TAS-Net by 0.07% in the metric P_{acc} on the Valence emotional potency binary classification task. The results on this dataset show that the proposed method is able to extract the EEG segments of subjects in emotionally strong states in an unsupervised manner very well.

Overall, our proposed method achieves better results on both SEED and DEAP datasets. Moreover, on the SEED dataset the Emotion-Agent extracts the key EEG segments that are more relevant to emotions in an unsupervised manner thus improving the performance of the results on the subsequent emotion recognition task. In addition, on the DEAP

dataset, for the task of emotion classification of Arousal labels, our method is based on our designed reward function, which measures the emotion intensity of EEG sample features and the emotion category to which they belong, and measures the interaction between EEG sample features and emotion prototypes well, and is able to accurately extract EEG emotion segments of people in emotionally intense moments in an unsupervised manner.

Discussion

To further validate our proposed Emotion-Agent model in terms of accuracy and reliability improvement in sentiment analysis, we conducted additional experiments on the SEED dataset. The experiments compared with and without, with i.e., using our proposed method Emotion-Agent extracts the key segments related to emotions and then inputs them into the classifier for the triple categorisation emotion recognition task, without on the other hand, we did not use our proposed method and input them directly into the classifier. On the other hand, we have chosen the traditional method SVM, KNN, and the deep learning supervised method MLP for the classifiers.

For emotion recognition with the SVM classifier, the accuracy P_{acc} reaches 56.30%, and P_f reaches 49.53%. After applying our proposed method, P_{acc} improves by 21.39% and P_f improves by 27.1%. With the KNN classifier, P_{acc} reaches 42.78%, and P_f reaches 34.93%. Following the implementation of our method, P_{acc} improves by 19.53% and P_f improves by 26.07%. For the MLP classifier, P_{acc} reaches 65.70%, and P_f reaches 63.55%. After using our proposed approach, P_{acc} improves by 12.02% and P_f improves by 13.21%. The experimental results show that the results on both traditional unsupervised, traditional supervised as well as deep learning supervised methods result in significant performance improvement on the SEED emotion recognition triple classification task. This experimental

Methods	Arousal		Valence	
	P_{acc}	P_f	P_{acc}	P_f
Supervised	Subject-Dependent		Video-Level LOOCV	
EMD (Zhuang et al. 2017)	71.99	-	69.10	-
Supervised	Subject-Independent		Subject-Level LOOCV	
DGCNN(Song et al. 2018)	61.10	-	59.29	-
ATDD-LSTM(Liang, Oba, and Ishii 2019)	72.97	-	69.06	-
Unsupervised	Subject-Independent		Subject-Level LOOCV	
EEG-FuseNet (Liang et al. 2021)	58.55	72.00	56.44	70.83
TAS-Net (Zhang et al. 2023)	60.51	72.64	57.84	71.80
Emotion-Agent(Ours)	75.06	73.88	57.91	55.43

Table 2: Model performance (%) for cross-subject emotion recognition on the DEAP dataset.

Methods	Sampling	Three-Class	
		P_{acc}	P_f
SVM	w / o	56.30	49.53
	w	77.69	76.63
KNN	w / o	42.78	34.93
	w	62.31	61.00
MLP	w / o	65.70	63.55
	w	77.72	76.76

Table 3: Emotion recognition performance (%) with the traditional classification methods using subject-independent LOOCV strategy on SEED dataset, under the conditions without (w/o) and with (w) the proposed Emotion-Agent method.

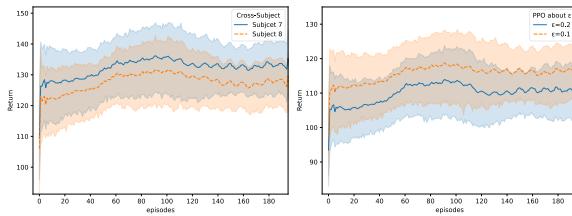


Figure 2: Evaluation on Optimization Process

result also shows that reward function we designed is representative of the intensity of human emotional states to a certain extent, and this reward function allows the Agent to autonomously and unsupervised select the key segment corresponding to the intensity of the emotion. This shows, to some extent, that our model extracts emotionally rich and relevant EEG segments, and then improves the accuracy and reliability of the downstream task analysis after feeding such relevant, partially informative emotional segments into the downstream task.

Model Optimization Process In order to have a better exploration of the EEG emotional space based on our well-designed reward function, we used PPO to complete the optimisation of the whole training process. To further study the role of the PPO algorithm in the model training process, we conducted additional experiments to explore the specific circumstances of the training process and the impact of the ϵ in the Clip operation on the model training process.

We conduct cross-subject experiments on the SEED dataset, and we compare the cumulative total return from the policy learned by the Agent during the completion of multiple Episodes of exploration and learning with the number of training sessions in the Cross-Subject experiments. Fig.2 depicts the alteration of the return during the cross-subject training of subjects 7 and 8 as the number of interaction episodes escalates. It can be observed that with a meticulously designed reward function, the cumulative benefits acquired by the agent in the task of extracting key EEG segments keep rising, and the agent progressively acquires the optimal action strategy. The right figure demonstrates the influence of the ϵ on the training process during training. It can be noted that when the upper limit of the clipping is lower, the training process is more stable and superior strategies are learned within a certain range. Constraining the ratio of new and old strategies enables the model to converge more steadily and efficiently during training.

Conclusion

In this paper, we propose a novel unsupervised deep reinforcement learning framework, called Emotion-Agent, to automatically identify relevant and informative emotional moments from continuous EEG signals. Emotion-Agent involves unsupervised deep reinforcement learning combined with a heuristic algorithm. Constructing heuristics for reinforcement learning by constructing prior knowledge for the exploration process can dramatically improve the efficiency of intelligences in the exploration process. The extraction of fragments that have a stronger connection to emotions is more favorable for the following analysis and research. Besides according to the results, we can show the effectiveness of our proposed approach.

References

- Apostolidis, E.; Adamantidou, E.; Metsai, A. I.; Mezaris, V.; and Patras, I. 2020. AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8): 3278–3292.
- Cheng, C.-A.; Kolobov, A.; and Swaminathan, A. 2021. Heuristic-guided reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 13550–13563.
- Cowen, A. S.; and Keltner, D. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38): E7900–E7909.
- Duan, R.-N.; Zhu, J.-Y.; and Lu, B.-L. 2013. Differential entropy feature for EEG-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, 81–84. IEEE.
- Gong, P.; Jia, Z.; Wang, P.; Zhou, Y.; and Zhang, D. 2023. ASTDF-Net: Attention-Based Spatial-Temporal Dual-Stream Fusion Network for EEG-Based Emotion Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 883–892.
- Hartung, J.; Knapp, G.; and Sinha, B. K. 2011. *Statistical meta-analysis with applications*. John Wiley & Sons.
- He, J.; Chen, J.; He, X.; Gao, J.; Li, L.; Deng, L.; and Oestendorf, M. 2016. Deep Reinforcement Learning with a Natural Language Action Space. arXiv:1511.04636.
- Horikawa, T.; Cowen, A. S.; Keltner, D.; and Kamitani, Y. 2020. The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. *Iscience*, 23(5).
- Huang, Y.-J.; Wu, C.-Y.; Wong, A. M.-K.; and Lin, B.-S. 2014. Novel active comb-shaped dry electrode for EEG measurement in hairy site. *IEEE Transactions on Biomedical Engineering*, 62(1): 256–263.
- Ivanov, D.; and Ben-Porat, O. 2024. Personalized Reinforcement Learning with a Budget of Policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12735–12743.
- Jiang, W.; Zhao, W. X.; Wang, J.; and Jiang, J. 2023. Continuous trajectory generation based on two-stage GAN. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4374–4382.
- Kalashnikov, D.; Irpan, A.; Pastor, P.; Ibarz, J.; Herzog, A.; Jang, E.; Quillen, D.; Holly, E.; Kalakrishnan, M.; Vanhoucke, V.; et al. 2018. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, 651–673. PMLR.
- Kletzander, L.; and Musliu, N. 2023. Large-state reinforcement learning for hyper-heuristics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12444–12452.
- Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1): 18–31.
- Li, J.; Qiu, S.; Du, C.; Wang, Y.; and He, H. 2019. Domain adaptation for EEG emotion recognition based on latent representation similarity. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2): 344–353.
- Li, R.; Wang, Y.; and Lu, B.-L. 2021. A multi-domain adaptive graph convolutional network for EEG-based emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5565–5573.
- Li, Y.; Zheng, W.; Cui, Z.; Zhang, T.; and Zong, Y. 2018. A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition. In *IJCAI*, 1561–1567.
- Liang, Z.; Oba, S.; and Ishii, S. 2019. An unsupervised EEG decoding system for human emotion recognition. *Neural Networks*, 116: 257–268.
- Liang, Z.; Zhou, R.; Zhang, L.; Li, L.; Huang, G.; Zhang, Z.; and Ishii, S. 2021. EEGFuseNet: Hybrid unsupervised deep feature characterization and fusion for high-dimensional EEG with an application to emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29: 1913–1925.
- Lin, Z.; Deng, Y.; and Pan, W. 2021. Combining the strengths of inverse-variance weighting and Egger regression in Mendelian randomization using a mixture of regressions model. *PLoS genetics*, 17(11): e1009922.
- Liu, Q.; Zhou, Z.; Wang, J.; and Liang, Z. 2024. Joint Contrastive Learning with Feature Alignment for Cross-Corpus EEG-based Emotion Recognition. *arXiv preprint arXiv:2404.09559*.
- Liu, Y.-J.; Yu, M.; Zhao, G.; Song, J.; Ge, Y.; and Shi, Y. 2017. Real-time movie-induced discrete emotion recognition from EEG signals. *IEEE Transactions on Affective Computing*, 9(4): 550–562.
- Martin, K. N.; and Arroyo, I. 2004. AgentX: Using reinforcement learning to improve the effectiveness of intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, 564–572. Springer.
- Ren, A. Z.; Govil, B.; Yang, T.-Y.; Narasimhan, K. R.; and Majumdar, A. 2023. Leveraging language for accelerated learning of tool manipulation. In *Conference on Robot Learning*, 1531–1541. PMLR.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Shi, L.-C.; Jiao, Y.-Y.; and Lu, B.-L. 2013. Differential entropy feature for EEG-based vigilance estimation. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6627–6630. IEEE.
- Song, T.; Zheng, W.; Song, P.; and Cui, Z. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541.
- Vecerik, M.; Hester, T.; Scholz, J.; Wang, F.; Pietquin, O.; Piot, B.; Heess, N.; Rothörl, T.; Lampe, T.; and Riedmiller, M. 2017. Leveraging demonstrations for deep reinforcement

learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *nature*, 575(7782): 350–354.

Warnell, G.; Waytowich, N.; Lawhern, V.; and Stone, P. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yarats, D.; Kostrikov, I.; and Fergus, R. 2021. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International conference on learning representations*.

Zhang, Y.; Pan, Y.; Zhang, Y.; Zhang, M.; Li, L.; Zhang, L.; Huang, G.; Su, L.; Liang, Z.; and Zhang, Z. 2023. Unsupervised time-aware sampling network with deep reinforcement learning for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*.

Zheng, W. 2016. Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis. *IEEE Transactions on Cognitive and Developmental Systems*, 9(3): 281–290.

Zheng, W.-L.; and Lu, B.-L. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3): 162–175.

Zhong, P.; Wang, D.; and Miao, C. 2020. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3): 1290–1301.

Zhou, K.; Qiao, Y.; and Xiang, T. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhou, R.; Zhang, Z.; Fu, H.; Zhang, L.; Li, L.; Huang, G.; Li, F.; Yang, X.; Dong, Y.; Zhang, Y.-T.; et al. 2023. PR-PL: A novel prototypical representation based pairwise learning framework for emotion recognition using EEG signals. *IEEE Transactions on Affective Computing*, 15(2): 657–670.

Zhuang, N.; Zeng, Y.; Tong, L.; Zhang, C.; Zhang, H.; and Yan, B. 2017. Emotion recognition from EEG signals using multidimensional information in EMD domain. *BioMed research international*, 2017(1): 8317357.

Zoph, B.; and Le, Q. V. 2017. Neural Architecture Search with Reinforcement Learning. *arXiv:1611.01578*.

Appendix

Datasets

We perform related experiments on the SEED (Zheng and Lu 2015) and the DEAP (Koelstra et al. 2011) using our proposed method Emotion-Agent. The following is a specific description of the two datasets:

SEED This dataset was developed by the BCMI laboratory at Shanghai Jiao Tong University. The dataset was acquired using the 62-channel ESI NeuroScan System based on the international 10-20 system, which recorded EEG signals from subjects under different types of video stimuli. The SEED dataset acquires raw EEG signals at a sampling rate of 1000 Hz. Regarding the experimental paradigm of the SEED dataset, specifically, the EEG signals of 15 subjects (7 males and 8 females) were recorded in the SEED dataset under various video stimuli. For each subject, the video clips to be viewed were divided into three different sessions. In each session, 15 different types of film clips were involved, among which there were three types of clips that elicited different emotional states (positive, neutral, and negative moods), and each emotional state comprised five film clips.

DEAP This dataset utilized a 32-channel Biosemi Active Two device with a sampling frequency of 512 Hz to record the subjects being stimulated by different one-minute-long music videos. Each video in the dataset corresponds to four labels, namely Valence, Arousal, Dominance, and Liking. A total of 32 participants in good physical condition were selected for the trial during the data collection process, consisting of 16 males and 16 females. Each subject was obligated to carry out 40 experiments, and in each of them, a 1-minute music video was watched to induce the relevant EEG. At the end of each experiment, a prompt self-assessment was conducted to rate the current state of the participant (Valence, Arousal, Dominance, and Liking), which was subsequently analysed and quantified comprehensively. Finally, a threshold value is employed to binarize the four labels for each video, thereby obtaining discrete labels for each state.

Table 1 shows the data statistics of the two datasets. We only use the preprocessed 1-s EEG signals from session 1 for both datasets. When conducting experiments with our model, we use the preprocessed 1-s EEG signals of session 1 for the SEED dataset. For the DEAP dataset, we likewise employ the preprocessed 1-s EEG signals.

Preprocessing

In the preprocessing section, we will respectively introduce the pre-processing of the two datasets and the extraction of the Differential Entropy Feature (DE feature) (Duan, Zhu, and Lu 2013) corresponding to the EEG signals.

The raw SEED dataset was initially preprocessed. To be specific, the raw EEG data were initially downsampled to a sampling rate of 200 Hz and filtered through a 1-75 Hz bandpass filter to filter out noise and eliminate artefacts. Next, the preprocessed EEG signals were divided into multiple segments by utilizing a sliding window with a length of 15 to obtain the EEG signals after the preliminary processing. In an effort to obtain features in the EEG signals that

are more closely related to the brain state, differential entropy features were extracted for the EEG signals measured in seconds (with a 200 Hz sampling rate corresponding to 200 sampling points) using a band-pass filter (δ wave 0.5-4 Hz, θ wave: 4-8 Hz, α wave: 8-13 Hz, β wave: 13-32 Hz, γ wave: 32-50 Hz). The specific expression for calculating the differential entropy of EEG signals is as follows:

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

where the time series X obeys the Gauss distribution $N(\mu, \sigma^2)$. It has been proven that, for a fixed length EEG sequence, DE is equivalent to the logarithm ES in a certain frequency band (Shi, Jiao, and Lu 2013). DE was employed to construct features in five frequency bands mentioned above.

After processing the EEG signal per second through differential entropy, the feature dimension changes from (62, 200) to (62, 5), where 62 represents the number of device channels. From the sample features per second, we extracted the features in 5 frequency bands of each channel and flattened them. Thus, the feature dimension of the sample features per second becomes (1, 310). We use the DE features of the EEG signal per second as the input of the model, which corresponds to three emotion labels (negative, neutral, and positive).

For the DEAP dataset, the original sampling frequency was 512 Hz. Subsequently, the data was downsampled to 128 Hz, while removing artefacts and deleting the first three seconds of silence in each experiment to obtain the initially processed EEG signal. Similarly for this dataset, to extract DE features (δ wave 0.5-4 Hz, θ wave: 4-8 Hz, α wave: 8-14 Hz, β wave: 14-32 Hz, γ wave: 32-50 Hz), the feature dimension per second changes from (32, 128) to (32, 5). We perform a flatten operation on it and the sample feature dimension per second is altered to (1, 160) to obtain the DE features of the EEG. We use the proposed model in Valence and Arousal two labels for experimentation, and both labels correspond to binary classification tasks.

Implementation Details

In this section, We will provide a detailed account of the specificities of the two processes, (1) Prototype Learning, and (2) Reinforcement Learning, that are employed in the model when training the model for cross-subject experiments (Subject-Leave-One-Out Cross-Validation). In order to better describe the training details, We define the total number of subjects in the dataset as N .

Prototype Learning

In the prototype learning stage, with the aim of obtaining a global overview of the data distribution, we cluster the data of $N - 1$ subjects through the utilization of the heuristic algorithm K-Means. In this case, the hyperparameter $n_clusters$ of K-Means was set to the total number of emotion categories (set to 3 for the SEED dataset and 2 for the DEAP dataset). After several iterations of the algorithm, we obtain the emotion prototype. Additionally, we compute the mean of the sum of squared errors in each cluster as the variance

of the data distribution within each cluster. We pass the sentiment prototypes μ_c for each emotion category and the variance σ^2 within each cluster to the second stage of learning.

Additionally, we employ the labels generated during the unsupervised clustering process as semantic information for subsequent EEG features, and thereby we define such a space as **Emotional Space**.

Reinforcement Learning

In the reinforcement learning stag, we require Trial-Level EEG data for delineation. The training data consist of the EEG DE features of a subject conducting an experiment to complete an indefinitely long sequence of actions in this manner as a decision-making process, where the action space is the discrete action $A = \{0, 1\}$.

The actor's network comprises one layer of LSTM where the number of hidden layer nodes is 128 and two fully connected layers where the hidden nodes range from $256 \rightarrow 128, 128 \rightarrow 2$. The network setup of the Critic is one layer of LSTM with 128 nodes and two fully connected layers where the hidden layer nodes in the fully connected layers are $256 \rightarrow 128, 128 \rightarrow 1$. The actor is optimized by Adam's optimizer and the learning rate is set to 1e-4, while the Critic's learning rate is also optimized by Adam's optimizer and is set to 1e-3. Additionally, in the PPO algorithm, γ is set to 0.98, λ is set to 0.95, and ϵ is set to 0.2 for the algorithms (refer to the original Eq. (9), Eq. (10), Eq. (11)). All experiments are carried out using PyTorch 1.13.1 on an NVIDIA GeForce RTX 3090 GPU.

Distribution-Prototype Reward

We obtain information about the global distribution based on the prototype learning stage. We use the mean and variance of each emotion cluster to reflect the individual variability within the distribution of EEG features for each emotion category. We propose two reward functions based on the distributional information derived from prototype learning: center reward and inter-intra reward.

We incorporate the relevant theory of **Inverse Variance Weighting** (Hartung, Knapp, and Sinha 2011) presented in Inter-Intra reward here.

If a series of independent measurements of a random variable are represented by y_i and possess a variance of σ_i^2 , then the inverse variance weighted average of these measurements is:

$$\hat{y} = \frac{\sum_i y_i / \sigma_i^2}{\sum_i 1 / \sigma_i^2}. \quad (16)$$

Among all the methods of weighted averaging, the inverse variance weighted average has the least variance. The expression of its variance is as follows:

$$D^2(\hat{y}) = \frac{1}{\sum_i 1 / \sigma_i^2}. \quad (17)$$

If the variances of the measurements are equalized, the inverse variance weighted average is the same as the simple average.

Datasets	Subject	Sessions	Trials	Channels	Sampling Rate (Hz)	Classes
SEED	15	3	15	62	1000	3(Negative, Neutral, Positive)
DEAP	32	1	40	32	512	2 Valence (Negative, Positive) 2 Arousal (Calm, Active)

Table 4: Detailed description of the SEED and DEAP datasets.

Additional Results

To further explore the EEG key segments extracted by our proposed method Emotion-Agent, we conduct additional experiments on the SEED dataset using the proposed model, visualized the extracted EEG keyfragments using t-SNE, and compared them with the case without.

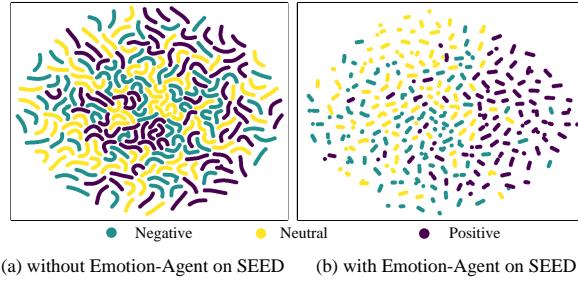


Figure 3: Comparsion of data distribution using t-SNE with and without Emotion-Agent on the SEED dataset.

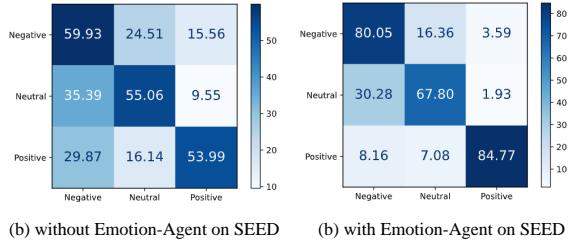


Figure 4: Comparsion of confusion matrices for emotion recognition with and without Emotion-Agent on the SEED dataset.

Figure 1 compares the with and without Emotion-Agent models, with SEED DE features as input, and visually compares the original DE features with the data after extracting key segments through t-SNE results. It can be seen that after extracting key segments, the separability of the entire data distribution is improved compared to before, and the same conclusion can be drawn from the improvement in classification accuracy.

Figure 2 compares the with and without Emotion-Agent models, with the SEED DE features as input and the classification confusion matrix obtained using SVM as the classifier. The experimental results show that compared with without, with achieves higher classification accuracy

on all three emotion categories.

EMOTION CLASSIFICATION FROM MULTI-CHANNEL EEG SIGNALS USING HiSTN: A HIERARCHICAL GRAPH-BASED SPATIAL-TEMPORAL APPROACH

A PREPRINT

 **Dongyang Kuang***

School of Mathematics (Zhuhai)
Sun Yat-sen University
Guangdong, 519082, CHINA
kuangdy@mail.sysu.edu.cn

 **Xinyue Song**

School of Mathematics (Zhuhai)
Sun Yat-sen University
Guangdong, 519082, CHINA
songxy39@mail2.sysu.edu.cn

 **Craig Michoski**

the Oden Institute for Computational Engineering and Sciences
University of Texas at Austin
201 E. 24th Street, POB 4.102
Austin, TX 78712, USA
michoski@oden.utexas.edu

August 29, 2024

ABSTRACT

This study introduces a parameter-efficient Hierarchical Spatial Temporal Network (HiSTN) specifically designed for the task of emotion classification using multi-channel electroencephalogram data. The network incorporates a graph hierarchy constructed from bottom-up at various abstraction levels, offering the dual advantages of enhanced task-relevant deep feature extraction and a lightweight design. The model's effectiveness is further amplified when used in conjunction with a proposed unique label smoothing method. Comprehensive benchmark experiments reveal that this combined approach yields high, balanced performance in terms of both quantitative and qualitative predictions. HiSTN, which has approximately 1,000 parameters, achieves mean F1 scores of 96.82% (valence) and 95.62% (arousal) in subject-dependent tests on the rarely-utilized 5-classification task problem from the DREAMER dataset. In the subject-independent settings, the same model yields mean F1 scores of 78.34% for valence and 81.59% for arousal. The adoption of the Sequential Top-2 Hit Rate (Seq2HR) metric highlights the significant enhancements in terms of the balance between model's quantitative and qualitative predictions achieved through our approach when compared to training with regular one-hot labels. These improvements surpass 50% in subject-dependent tasks and 30% in subject-independent tasks. The study also includes relevant ablation studies and case explorations to further elucidate the workings of the proposed model and enhance its interpretability.

Keywords Affective Computing · Emotion Recognition · EEG · Hierarchical Spatial Temporal Network · Parameter Efficient Models

*<https://github.com/dykuang>

1 Introduction

Initially proposed in Picard [2000], the field of affective computing has since evolved to play a significant role within artificial intelligence. Among the multitude of data sources leveraged to discern human psychological states, non-invasive electroencephalogram (EEG) stands out due to its various advantages. These include but are not limited to its portability, relatively high temporal resolution, and assured safety. Integrating EEG-based affective computing with an array of pattern recognition tools, particularly the rapidly evolving neural network methodologies in deep learning, exhibits substantial potential across a multitude of applications Gong et al. [2021], Wang et al. [2022], Li et al. [2022d]. The preliminary step in comprehending human emotions is the construction of models that admit quantifiable parametric relationships.

Fundamentally, two abductive categories of models are usually identified to exist for quantifying emotions; the first, termed *discrete quantification models*, are often cited as being scientifically grounded in the early work of Ekman Ekman [2009] and Plutchik Plutchik [2003], though are rooted in the work of the ancient philosophers, such as Aristotle Aristotle [1984], Seneca Seneca [2017], and Epictetus Epictetus [2004]. These models envision the emotion space as ‘patches’, each representing a basic state such as, e.g., anger, anticipation, fear, sadness, disgust, trust, surprise, joy, etc., as well as their various combinations and embodiments. The second category of models (which might be viewed as a refinement of the first allowing for partial inclusions and multidimensional amalgams), are the *dimensional quantification models*, which employ mutually orthogonal axes to construct distinct (or independent) emotional dimensions. For example, within the realm of affective computing, Russell’s Valence-Arousal bipolar emotional quadrant system Russell [1979] has gained wide acceptance, where the Valence axis aids in gauging an individual’s happiness or sadness, while the Arousal axis quantifies the level of excitement. These geometry rich relations described by different emotion-based models provide many challenges for prediction frameworks in regular classification problems where OneHot labels are used to seek a model whose predictive behavior/logic is closer to that of human beings. A particular aspect of this concern is discussed more in part B of Section 2.

Within the realm of numerical frameworks, recent successes of large-scale deep learning models, such as ChatGPT, have generated considerable interest in both industry and academia. However, in contrast to models such as ChatGPT, where huge datasets are available to train on, in the field of EEG-based human emotion recognition, a significant barrier impeding the effectiveness of large models in applications is data limitation. Despite researchers contributing to open-source datasets like SEEDZheng and Lu [2015], DEAPKoelstra et al. [2011], DREAMERKatsigianis and Ramzan [2017], ASCERTAINSubramanian et al. [2016], etc., the overall volume of available data remains **extremely limited**, introducing a challenging problem when attempting to develop robust predictive and analytical frameworks. Nevertheless, extensive work on deep learning models in EEG-based emotion recognition has been done involving nearly all mainstream types of neural networks, including CNNsLawhern et al. [2018], Wu et al. [2022], RNNs/LSTMsTao et al. [2020], Cui et al. [2020], capsule networksLi et al. [2022a], Wei et al. [2023], graph convolution networks (GCNs)Song et al. [2021], Priyasad et al. [2022], Liu et al. [2022], Li et al. [2022c], transformersWei et al. [2023], etc. In addition to these network models, active areas of study in the domain of EEG signal analysis also include: attention module designs Kuang et al. [2023], model compressing techniques Liu et al. [2022], and domain transfer learning approaches Quan et al. [2023]. It is, however, also worth pointing out that prediction tasks for fine-grained labeling approaches (e.g. valence level 1,2,...,5) are significantly less studied/understand than those of the more standard binary classification tasks (e.g. high v.s. low valence) commonly seen in benchmark studies.

In addition to the frequent pursuit of enhanced predictive accuracy across diverse tasks, relatively fewer recent studies have concentrated on the qualitative characteristics of learning-based, data-driven models within the specific context of EEG-based emotion recognition. However, along these lines, in the pioneering work of Lawhern et al. [2018] a compact model is proposed for accommodating the limited-data concern. In this work the spatial-temporal nature of the signals is considered, but it is addressed via a simple approach by convolutional operations with kernels of custom sizes, which is more of a technique arising from general practices previously found within the computer vision community (note: in that context, though, without involving or adapting to EEG priors). Additionally, works like Wu et al. [2022] further advance on the special spatial-temporal nature of EEGs by introducing a Multi-Scales Bi-hemispheric Asymmetric Model (MSBAM) that recasts the original EEG signal format into a sparse matrix representation of 3D input features by exploring a network design that incorporates the brain’s bi-hemisphere asymmetry. The resulting MSBAM design, however, leads to a much larger network and the information learned from the “empty” entries in the constructed sparse input is still reliant on a significantly large latent space (i.e. conventional “black-boxes”). Comparatively, an example in work focused on using graph convolutional networks utilizes the technique of DiffPool Ying et al. [2018], allowing for automatic graph hierarchies—though this approach was not originally intended for use on time series inputs. Because of this, notable augmentations and adaptations must be adopted to incorporate emotion-based EEG priors for enhancing the model performance. It is, however, also worth mentioning that none of these previous studies consider the question

of whether the learned model’s predictions (e.g. feature representation space) are inherently consistent between the different labeling models (i.e. for example, the aforementioned discrete and multidimensional emotion labeling models).

Drawing inspiration from prior research and identifying opportunities for enhancement within their methodologies, we introduce a Hierarchical Spatial Temporal Network (HiSTN) design. This approach seeks to harmonize the objectives of a lightweight model architecture with the establishment of a temporal-spatial hierarchy that integrates interpretable priors, while also ensuring that the model’s predictive behavior aligns more closely with established clinical models of emotion. The key contributions of this paper are delineated as follows:

1. We propose a lightweight, parameter-efficient design tailored for prediction tasks that have limited training data.
2. We incorporate a hierarchical graph convolution component to extract spatial-temporal features at varying levels of abstraction. This design facilitates an intuitive interface for integrating prior knowledge about potentially useful hierarchical spatial information.
3. We introduce a special label smoothing technique that enhances the model’s qualitative behavior, particularly in terms of continuity among the model’s highest-ranked predictions. This technique helps to ‘shape’ the learned feature representation space for tasks with categorical labels from multidimensional emotion models.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive explanation of our ideas for the network and training design, including the motivations behind these decisions from related work and a special label encoding method proposed for better consistency between numerical categorical labels and clinical emotion models during learning. Section 3 presents our main benchmark results for both subject-dependent and -independent tasks. This section also includes relevant case studies and ablation studies that investigate the effects of some crucial choices made during model specification and training strategies.

2 The Motivation and Idea

2.1 Network Design

Generally speaking, EEG signals recorded from non-invasive devices have relatively good temporal resolution. The regular time convolution layers, such as a one dimensional convolution layer, tend to be quite effective for extracting time related features hierarchically when being properly stacked in a network design framework. On the other hand, spatial information is often considered lacking in non-invasive EEG signals. Due to the limitation in EEG signals, and the considerable utility and value of spatial information, much research has been performed aimed at enhancing the relatively poor spatial resolution of EEG signals, and then designing these enhancements into the network construction process itself.

For example, prior studies, exemplified by Song et al. [2021], Li et al. [2022b], have utilized diverse projection and interpolation techniques to convert multi-channel time series data into a format compatible with image-based data. In a similar vein, Wu et al. [2022] introduced a method that arranges distinct channels into a sparse 2D matrix representation, in order to reveal certain spatial signal components. However, the emergence and growing acceptance of graph convolutional neural networks (GCNs) have prompted researchers to adopt this framework with increasing regularity when developing spatially resolving models.

There are at least *two crucial issues* that arise from the GCN approach. The first pertains to *the construction of an appropriate graph*. Previous practices encompassed the use of manually specified graphs based on clinical priors Wang et al. [2019], Tian et al. [2022], employing different functional connectivity measures, or dynamically generating/adjusting them during the learning process Song et al. [2018, 2021], Priyasad et al. [2022]. The second key concern arises during the graph embedding step. Traditionally, this step involves Multi-Linear Perceptron (MLP) layers that map node features at layer $l - 1$ with dimension d_{l-1} to the desired node feature dimension d_l at layer l (where d_{l-1} and d_l may be equal) in the subsequent layer. However, if one tries to utilize GCN with signals directly as node features, as determined by the sampling frequency and time window, the GCN layer would accumulate *a significant number of parameters* during this embedding process. To mitigate this, related work either conducts manual feature extractions beforehand and employs them as input instead of the original signal Priyasad et al. [2022], Li et al. [2022b], or positions the graph layer in deeper blocks of the network design, where the time dimension of learned abstract features is reduced by previous time convolutional blocks Song et al. [2021]. Our work in this paper aligns with these ideas, but with more focuses on a light-weighted yet structure-rich graph design.

Our overall design, as depicted in Fig. 1, consists of three main parts: a ‘**Feature Head**’ for successive feature preprocessing and auto-extraction (I), a ‘**Hierarchy Core**’ for developing information hierarchy via proper graphs

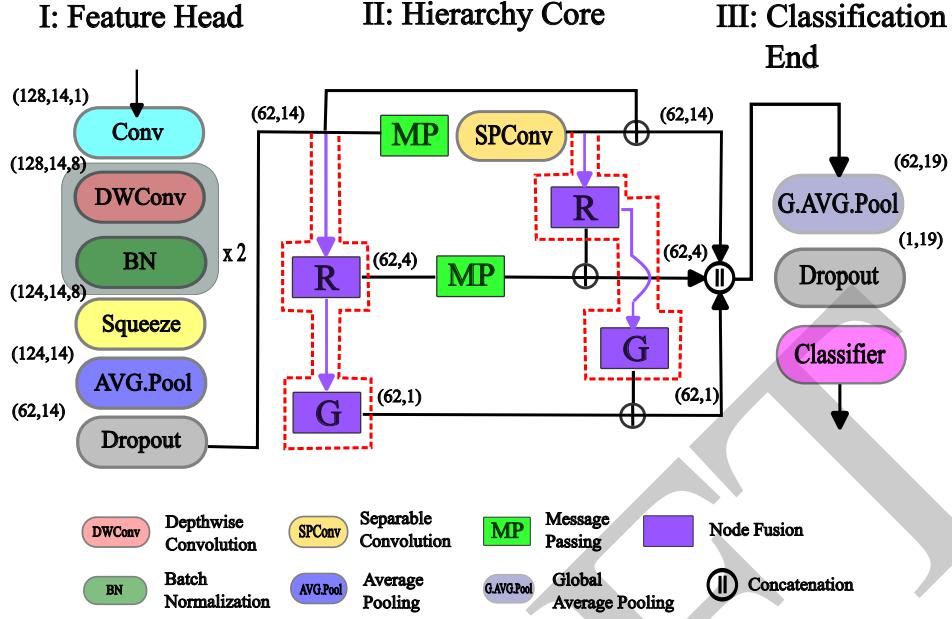


Figure 1: The design of the proposed HiSTN network. A closer look at the Hierarchy Core (enclosed by the red dashed line) is unpacked in Fig. 2.

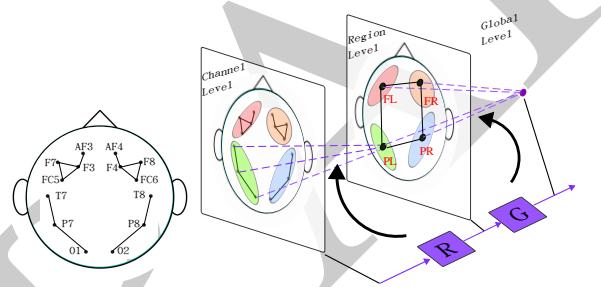


Figure 2: Unpacking the node fusion block. FL: Frontal Left, FR: Frontal Right, PL: Parietal Left, PR: Parietal Right. At the intermediate stage, Region block R “summarizes” the learned information from previous node/channel level features. This processed information per region is then further summarized by the Global block G over the whole graph.

(II) and a ‘Classification End’ (III). In line with previous research like EEGNet Lawhern et al. [2018], the ‘Feature Head’ comprises a series of stacked temporal convolution layers for extracting lower-level features primarily focused on temporal aspects. This is achieved by using convolution kernels of shape $(k, 1)$, $k > 1$ (in all convolution layers) limiting the convolution to the time dimension. It is also worth noting that the very first 2D convolution layer expands a typical spatial-temporal signal of shape (T, C) to (T, C, S) by learning S different convolution kernels that provide the “multiple view:” $X^m \in \mathcal{R}^{T \times C \times S}$ for further feature extraction. That is, the squeeze layer learns proper weights for summarizing previous multiple views into one via setting $X^{squeezed} = \sum_{i=1}^S w_i X_i^m$ where X_i^m denotes the i th view. This also facilitates dimension reduction along the time axis via temporal pooling, resulting in reduced computational complexity during the graph message passing stage in Part II. The ‘Hierarchical Core’ enriches the hierarchical spatial information by establishing different levels of graphs (channel level, region level, and global level) through node fusion blocks. Message passing is performed on the corresponding graphs at each level. The nodes at the channel level correspond to individual EEG channels, while the nodes at the region level and graph level are abstract concept nodes. The information flow within this block will be introduced below. Finally, Part III is a standard ‘Classification End’ consisting of pooling, dropout layers, and a relevant classifier.

Fig. 2 offers insight into the pathway (highlighted in purple and enclosed by the red dashed line in Fig. 1) responsible for constructing a graph hierarchy. An example is presented by partitioning all channels into four regions. The graph formed by AF3-F3-F7-FC5 is denoted as FL, representing the Frontal Left region. Similarly, PL represents T7-P7-O1, encompassing the temporal/parietal/occipital left region. FR and PR are defined similarly for regions on the right hemisphere. This hierarchical graph structure is illustrated in Fig. 1, with tensor shapes marked to aid readers in

tracking the flow of information. For instance, the node fusion block labeled with the letter ‘‘R’’ is responsible for fusing channel-level feature of shape (62, 14) into region-level features of shape (62, 4). The fusion block, denoted as ‘‘G’’, further abstracts region-level features to form graph-level features of shape (62, 1). These hierarchical deep features learned at different spatial levels are then concatenated resulting in a feature tensor of shape (62, 19), which is then globally pooled along time dimension forming a summary feature vector of length 19 before passing to the classifier in Part III.

Information flow occurs between different nodes based on the underlying graph structure at each level, with information being transmitted or fused exclusively from lower levels to higher levels to emulate the process of abstraction. While a fully connected graph can be used at the channel level, we opt for a graph comprising N connected components for a couple reasons. First, it prevents potential interference from nodes belonging to other regions during feature fusion at the next level. Second, it simplifies the graph structure and facilitates future parallel implementation. This design offers a convenient interface for incorporating specific prior knowledge about functional connectivity graphs. Further exploration of graph choices is conducted in Section 3.4 through ablation studies.

The feature fusion from the previous to the subsequent layer (purple dashed line in Fig. 2) can be described as follows:

$$X_i^{(l)} = \sum_{j \in N_i^{(l-1)}} C_j^i X_j^{(l-1)}, \quad C^i = \varphi \left(\text{MLP}(X_i^{(l-1)}) \right), \quad (1)$$

such that $X_i^{(l)}$ is the node i ’s feature at level l formed by the weighted sum from its corresponding node features $X_j^{(l-1)}$ at level $l - 1$. The collection of node i ’s neighboring nodes at the previous level $l - 1$ is denoted by $N_i^{(l-1)}$. The learnable weight vector $C^i = [C_1^i, \dots, C_{N_i}^i]$ correspond to the outputs from a single MLP layer activated by the regular softmax function φ . As for the graph message passing within each graph convolution level, we adopt the following method using a Chebyshev polynomial on the normalized graph Laplacian matrix as the graph convolution kernel for faster and more stable approximations Defferrard et al. [2016], He et al. [2022]:

$$\hat{X}_i^{(l)} = \sigma \left(\sum_{k=0}^d \beta_k T_k(\tilde{L}) X_i^{(l)} \right), \quad (2)$$

$$\text{where } L = D - A, \quad \tilde{L} = \frac{2L}{\lambda_{\max}} - I. \quad (3)$$

Here, σ is a nonlinear activation function, A is the graph adjacency matrix, D is the corresponding diagonal matrix with diagonal entries being the degree of the corresponding node, and λ_{\max} is the maximum eigenvalue of L . The k -th order Chebyshev polynomial can be obtained conveniently via the usual recurrence formula:

$$T_k(\tilde{L}) = 2\tilde{L}T_{k-1}(\tilde{L}) - T_{k-2}(\tilde{L}), \quad (4)$$

$$\text{where } T_0(\tilde{L}) = I \text{ and } T_1(\tilde{L}) = \tilde{L}. \quad (5)$$

In our experiments in Section 3, the maximum degree d is chosen to be the diameter for each graph considered. As for the feature processing after message passing, we use a 1D separable convolution along the feature’s time dimension instead of dense layers for capturing possible time related information in the nodal features. With the adopted graph shown in Fig. 1, features learned from these three levels are later concatenated together. The concatenated feature is then globally averaged/pooled along the time dimension before being fed to the classifier. It is important to highlight key distinctions between our approach and other works employing graph hierarchies, such as DiffPool Ying et al. [2018], which focuses on general and automatic graph pooling for hierarchical learning. DiffPool initiates with a connected graph and dynamically learns assignment matrices for node fusion, adjacency matrices for message passing, and feature embedding matrices during training. Each node at level $l - 1$ can be associated with all nodes at the next level, which differs from our aforementioned design. Additionally, instead of regular graph embedding via matrix multiplication, we utilize a separable convolution along the time dimension for feature embedding, considering that the features at each node are time series features. Time convolution (\mathcal{T})² requires fewer parameters and is generally faster than MLP. However, it does not commute with the message passing operation (ψ) mathematically in general, i.e. $\mathcal{T} \circ \psi \neq \psi \circ \mathcal{T}$, prompting us to adopt the multi-branched design in Part II of Fig. 1 for feature fusion before and after the aforementioned operations.

2.2 Prediction Framework

One of the goals in current and future deep learning tools is to make prediction models more human-like, or more capable of emulating clinical human responses or researcher-compatible ‘interpretive evaluations.’ For example, in the

²This time convolution is performed as depthwise convolution in implementation.

context of training models for EEG-based emotion recognition, especially for score-based predictions, given a particular dataset/context with a fixed emotional stimulus ‘type’, there often exists certain levels of uncertainty/fuzziness in the self-assessed rating. For instance, if someone’s self-assessment during a emotional stimuli type is 6 on a 1-9 scale, the ‘subjectivity’ of personal perception, together with the natural order of the scale, implies nearby scores of 5 or 7 are more likely when tested at a different time than more extreme variations such as a score of a 1 or a 9. A standard OneHot labeling approach does not address these logical priors well and it can lead to an over-confidence concern when training a model Guo et al. [2017]. As an easy and straightforward approach for incorporating this logical prior into the model, which we refer to as the “Continuum of Predictions (CoP)” behavior for trained models, a prior distribution centered at the maximum likelihood of the existing self-assessed scores can be utilized for modifying training labels. While previous prediction methods have primarily focused on top-1 accuracy or similar performance metrics, the consideration of this “Continuum of Predictions” has received less attention, especially under fine-grained predictions which we adopted for model comparisons in this paper. It is worth noting here that while loss functions such as mean squared error or mean absolute error, instead of the widely adopted cross-entropy loss for classification tasks, can naturally address the prediction continuum problem, research in the field of “learning with noisy labels” has shown that these loss functions’ generalization performance significantly degrades when dealing with complex data, compared to categorical cross-entropy (CCE) loss Ghosh et al. [2017], Song et al. [2022]. Thus, as a second objective in this paper, we present an easy variation of the classical label smoothing technique Szegedy et al. [2016], Müller et al. [2019] to address the issue of CoP under the task of fined-grained emotion score predictions. To enable better comparison, we explore four different classifier designs in this work.

- A:** The model is trained using regular OneHot label encoding, where the label 2 is encoded as $\{0, 0, 1, 0, 0\}$, and trained with Categorical Cross-Entropy (CCE) loss between true labels y and the associated predictions \hat{y} :

$$L(y, \hat{y}) = - \sum_{i=1}^N y_i \log(\hat{y}_i).$$

- B:** The model’s output directly predicts the subject’s self-reported score, and it is trained using mean absolute error (MAE) loss: $L(y, \hat{y}) = - \sum_{i=1}^N |y_i - \hat{y}_i|$.

- C:** The model’s output consists of the parameters θ describing a Gaussian Mixture Model (GMM) with 5 components ($N = 5$), and it is trained using Negative Log-likelihood Loss, $L(y, \hat{y}) = - \log(p_{\text{GMM}}(\hat{y}; y; \theta))$, based on the GMM probabilities.

- D:** The model is trained using a specially smoothed label encoding method (see Eq-6 below). For example,

$$\{0, 0, 1, 0, 0\} \rightarrow \{2.64 \times 10^{-4}, 0.11, 0.79, 0.11, 2.64 \times 10^{-4}\}$$

represents the smoothed encoded label for score 2. The model is also trained with Categorical Cross-Entropy (CCE) loss. This is the proposed label encoding method.

Eq-6 gives the formula smoothing the label for addressing the prediction ‘continuum’ problem, where i is the true label and j is the index corresponding to the j -th score in the smoothed label. The P_j gives the smoothed value³ at index j , while $s = 0.5$ (half the width between two consecutive ratings) is used in our study here, yielding:

$$P_j = \frac{\exp(-(j-i)^2/2s^2)}{\sum_j \exp(-(j-i)^2/2s^2)}. \quad (6)$$

Our benchmark experiments in Section 3 have shown that this simple modification of label encoding can greatly help improve the resulting model’s top predictions. Codes relevant with this paper will be made available on Github at <https://github.com/dykuang/EEG-based-affective-computing>.

3 Experiment

3.1 The DREAMER Dataset

The DREAMER Dataset Katsigiannis and Ramzan [2017] is a multimodal database containing EEG and ECG signals recorded during the elicitation of affect using audio-visual stimuli. The dataset comprises data from 23 subjects (14 males and 9 females), including their self-assessments (integers from 1 to 5) in terms of valence, arousal, and dominance after each stimulus. The dataset consists of 18 film clips, with varying durations (ranging from 65 seconds to 393 seconds), and for detailed information about each film clip, readers can refer to Gabert-Quillen et al. [2015]. It is worth

³It can also be viewed as a fuzzy membership value on the label set.

noting that some subjects' labels do not cover the full range of scores for arousal and dominance, but only a subset of them. For the EEG signal collection, the Emotive EPOC wireless headset and the Shimmer 2 ECG sensors were utilized.

In this particular experiment, we solely utilized the EEG signals, which consists of a total of 14 channels, namely AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4. Previous research studies have demonstrated promising results in binary classification tasks using this dataset, considering both the subject-dependent and the subject-independent settings (refer to Table 1). For binary classifications, a threshold (such as 3) is selected to map the original 5 scores into 2 classes (high vs. low). The works presented in Table 1 can differ in some details about experimental settings, including training data preparation, normalization methods, and evaluation criteria, among others.

Table 1: A collection of previous and recent work on the binary classification task using DREAMER. The Percentage refers to accuracy (mean \pm std.dev.). Rows corresponding to subject-independent experiments are colored in cyan. The Notes column collects the evaluation method and length of signals used for prediction.

Models	Year	Valence(%)	Arousal(%)	Notes
GECNN Song et al. [2021]	2021	95.73 \pm -	92.79 \pm -	leave one trial out CV, 2s
DCNN+GAT-MHA Priyasad et al. [2022]	2022	88.80 \pm -	88.24 \pm -	10CV, 1s
SFC SANLi et al. [2022b]	2022	93.77 \pm -	95.80 \pm -	10CV, 1s
MTCA-CapsNet Li et al. [2022a]	2022	94.96 \pm 3.60	95.54 \pm 3.63	10CV, 1s
ACRNNTao et al. [2020]	2022	97.93 \pm 1.73	97.98 \pm 1.92	10CV, 1s
RGCBLi et al. [2022c]	2022	87.43 \pm 14.89	91.55 \pm 14.78	leave one session out CV, 1s
Bi-CapsNet Liu et al. [2022]	2023	95.48 \pm 3.26	95.86 \pm 3.31	10CV, 1s
TC-Net Wei et al. [2023]	2023	98.59 \pm 1.38	98.61 \pm 1.34	10CV, 1s
MSBAM Wu et al. [2022]	2023	99.69 \pm 0.24	99.76 \pm 0.20	10CV, 1s
TDMNN Ju et al. [2023]	2023	99.45 \pm 0.51	99.51 \pm 0.79	5CV, 3s
RM CNN Maheshwari et al. [2021]	2021	58.02 \pm -	51.23 \pm -	LOOCV, 10s
SparseD Zhang et al. [2021]	2021	64.06 \pm 8.58	66.96 \pm 6.91	LOOCV, 2s
FLDNet Wang et al. [2021]	2021	89.91 \pm 12.51	87.67 \pm 10.02	5CV, 3s
DCNN+GAT-MHA Priyasad et al. [2022]	2022	64.98 \pm -	63.71 \pm -	LOOCV, 1s
MR-VAE-DT Quan et al. [2023]	2023	73.08 \pm 8.84	72.74 \pm 12.93	LOOCV, 1s
DCNN+NN Singh et al. [2023]	2023	96.74 \pm -	97.64 \pm -	80/10/10 split, 1s

This paper focuses on the more challenging and less studied task of 5-classification for better assessing not only the proposed network's learning ability, but also the qualitative improvement our proposed label smoothing trick can bring to the prediction. To account for the time it can take for emotions to develop Lerner et al. [2015], Adolphs [2002], not all data is utilized in our benchmark evaluations. Following previous research as in Table 1, we consider the last 1 minute of data during the stimuli for our experiments. The data is normalized using the corresponding 1-minute baseline signal when no stimuli are presented. In our benchmarks, we also include EEGNet as a representative compact model baseline and MSBAM as an example of a larger network utilizing matrix embedding and bi-hemisphere asymmetrical priors to enhance spatial information for comparison.

The F1 score is utilized to evaluate the top prediction performance across different configurations. Additionally, the top-2 accuracy is employed to further compare the predictive capabilities of trained models. In addition to commonly used quantitative measures, we introduce two additional metrics for assessing the qualitative behavior of the models in terms of the mentioned "Continuum of Prediction". The first metric, referred to as "Tridiagonal Percentage (Tri-P)", is calculated as $\text{Tri-P} = 100 \times \left(\frac{\sum_{|i-j|<2} C_{ij}}{\sum_{i,j} C_{ij}} \right)$ where C represents the confusion matrix, and this quantity represents a percentage ratio between the sum of elements that are on the diagonal, subdiagonal, or superdiagonal and the number of total test samples. The higher the metric is, the better the overall prediction quality is in terms of CoP. The second metric measures the proportion of samples in the test set that meet two conditions: (1) the model's top-2 predictions

for the sample are consecutive, and (2) the true label for the sample falls within the model’s top-2 predictions. In this paper, we refer to this metric as the “Sequential Top-2 Hit Rate” or ‘Seq2HR’. A higher value indicates better overall prediction performance. Within each benchmarked model, we will compare the four variations with different classifier designs as mentioned in Section 2. All numerical experiments are performed with *Tensorflow* framework on a *Nvidia 3080Ti* graphic card.

3.2 Subject-Dependent Experiments

Table 2: Hyper-parameters used in benchmark.

Batchsize	256
LearningRate	0.001
Maximum Epochs	50

As mentioned earlier, even if some subjects’ self-reported labels contain only a subset of the 5 ratings, we maintain consistency in predictions among all subjects by using the same model configuration with 5 output classes. A special 10-fold cross-validation (10 CV) approach is employed for the benchmarks in this section. The 1-minute data is divided into 10 consecutive non-overlapping trunks, each lasting 6 seconds. During each iteration, one fold is used for testing, the fold preceding it is used for validation, and the remaining folds are used for training. To address data limitations and potential label imbalance issues, a data generator is utilized to randomly select 1-second segments from the training data. Within each batch, an equal number of samples is associated with each different label. The model with the highest validation performance is saved,

and its performance is evaluated on the test set, which consists of 1000 randomly generated 1-second samples from the mentioned data generator. The Adam optimizer is utilized for training, and additional important hyperparameter settings are detailed in Table 2. The benchmark performance for the 5-class prediction, under different model configurations, is summarized in Table 3. The best performance per metric column overall is highlighted in **bold**, while the best performance for each model among different training variations is colored with an orange background. The second-best performance for each model is colored with a cyan background.

Table 3: Summary of subject-dependent experiments on 5-classification tasks.

		Valence				Arousal				
Models	#Para.	F1(%)	Top2 Acc.(%)	Tri-P(%)	Seq2HR (%)	F1(%)	Top2 Acc.(%)	Tri-P(%)	Seq2HR (%)	
EEGNet	A	3269	94.74 ± 3.82	98.98 ± 0.81	97.09 ± 2.35	42.80 ± 13.54	95.59 ± 3.40	99.01 ± 1.06	98.28 ± 1.71	52.19 ± 18.07
	B	1217	63.22 ± 14.31	86.10 ± 9.37	93.57 ± 4.76	86.10 ± 9.36	67.29 ± 17.24	89.53 ± 9.30	95.48 ± 5.44	88.63 ± 10.22
	C	8399	95.10 ± 4.16	98.78 ± 1.33	98.00 ± 1.88	59.48 ± 15.39	96.02 ± 3.22	99.18 ± 0.64	98.54 ± 1.43	62.86 ± 18.11
	D	3269	91.13 ± 6.00	92.44 ± 2.41	96.69 ± 3.16	91.71 ± 6.30	90.95 ± 5.80	97.76 ± 1.81	97.07 ± 2.36	92.72 ± 5.58
MSBAM	A	203,269	92.70 ± 3.96	97.70 ± 1.61	95.73 ± 2.55	41.46 ± 13.94	92.90 ± 4.76	98.06 ± 1.85	97.28 ± 2.47	54.36 ± 19.35
	B	202,785	88.49 ± 6.06	96.15 ± 2.51	98.07 ± 1.34	96.15 ± 2.51	89.60 ± 7.44	96.95 ± 3.02	98.62 ± 1.61	96.93 ± 3.02
	C	204,479	95.77 ± 2.49	98.29 ± 1.23	98.03 ± 1.36	72.10 ± 10.53	95.26 ± 3.03	98.30 ± 1.20	98.46 ± 1.37	64.42 ± 13.71
	D	203,269	95.29 ± 3.01	98.26 ± 1.28	98.16 ± 1.53	96.11 ± 3.16	95.63 ± 2.68	98.47 ± 1.34	98.71 ± 1.29	97.51 ± 2.35
HiSTN	A	1181	97.13 ± 1.60	99.59 ± 0.29	98.49 ± 0.95	45.22 ± 12.88	97.33 ± 2.20	99.67 ± 0.45	98.97 ± 1.43	47.72 ± 19.46
	B	1101	82.08 ± 6.48	95.98 ± 1.99	98.41 ± 0.98	95.98 ± 1.98	85.61 ± 8.74	96.73 ± 3.53	98.82 ± 1.59	96.70 ± 3.53
	C	1381	94.13 ± 3.16	98.77 ± 1.00	98.42 ± 0.97	75.80 ± 11.52	94.53 ± 4.51	98.81 ± 1.22	98.49 ± 1.57	69.34 ± 12.25
	D	1181	96.82 ± 1.65	99.28 ± 0.53	99.22 ± 0.72	97.57 ± 2.15	95.62 ± 2.96	99.23 ± 0.67	99.29 ± 0.80	97.82 ± 1.96

Among the three compared network designs, MSBAM has the highest number of parameters (approximately 200k) among the selected architectures, but it is still not considered large compared to most other methods listed in Table 1. While EEGNet is already designed to be compact, the proposed HiSTN used in this experiment significantly reduces the parameters to only about **1k**. In terms of F1 score, HiSTN-A (HiSTN with regular OneHot encoding) achieves the highest values for both valence (**97.13%**) and arousal (**97.33%**) prediction. HiSTN-D (HiSTN with specially smoothed label encoding) outperforms other configurations in terms of other metrics, particularly showing notable improvements (at least **50%**) for Seq2HR. Similar improvements are observed with EEGNet and MSBAM, indicating the universal effectiveness of enhancing the prediction’s “continuum” behavior⁴.

Although conducting classification as a regression task (training type B) ensures perfect prediction “continuum” behavior, the distance-based loss is not as effective as the softmax-typed loss in training the model for accuracy. For all three models considered, the corresponding F1 scores and Top2 accuracy are the lowest among the four training variations compared. This suboptimal accuracy behavior also impacts other metrics such as Tri-P and Seq2HR. When

⁴Pairwise t-test gives very small p-values (i.e. EEGNet: $4.92e^{-14}/4.09e^{-11}$, MSBAM: $4.28e^{-14}/2.03e^{-11}$, HiSTN: $2.25e^{-14}/1.59e^{-11}$, for valence/arousal respectively, suggesting strong statistical significance) when testing if the proposed special label smoothing can bring improvements to Seq2HR against regular OneHot label encoding.

comparing MSBAM and HiSTN with EEGNet, notable parameter differences can be observed between training variations B and C, primarily influenced by the transition from flattened features to the last dense layer. The number of parameters is sensitive to the hidden units in the last layer, with EEGNet exhibiting this sensitivity more prominently. MSBAM and HiSTN, on the other hand, demonstrate a more balanced distribution of parameters across shallow and deep layers, resulting in reduced sensitivity to the hidden units in the last dense layer. While training variation C (model outputting a mixed Gaussian distribution) shows improved Seq2HR despite having the most parameters, it remains competitive across other metrics. In the experiments with EEGNet, it achieves the highest F1 scores and Top2 accuracy for both valence and arousal score prediction tasks. For experiments with MSBAM, it achieves the highest F1 scores and Top2 accuracy in the valence score prediction task and the second highest (very close to the highest value) in the arousal score prediction task.

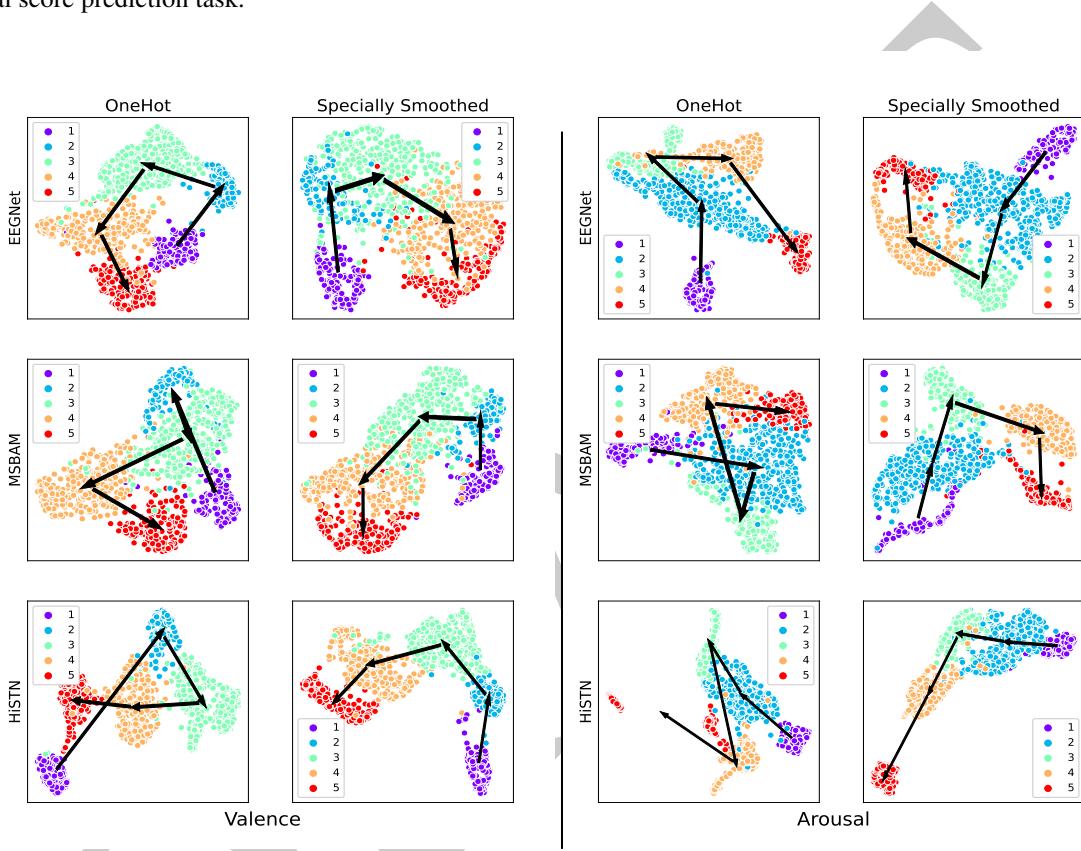


Figure 3: 2D embedding of deep features using UMAP when different models are trained with (1) regular OneHot label encoding, and (2) our proposed special label-smoothing. Left: Valence prediction with data from Subject S3, Right: Valence prediction with data from Subject S23.

3.2.1 Case Study – Representation Space

As discussed and validated in Table 3, training with our proposed specially smoothed labels can significantly help increase the prediction quality in terms of CoP. To provide a different visualization perspective for gauging the improvements these smoothed labels can bring to the learned feature representation space, we also present Fig. 3. In this figure, features output by the last dense layer before activation are embedded into two dimensions via the UMAP algorithm Sainburg et al. [2021]. In addition, arrows from the cluster center ranked i to cluster centers ranked $i + 1$ are also attached, as further shown in Fig. 3. From this plot, one can easily see that the manifold bearing representations belonging to different ranking scores ($1 \rightarrow 5$) is more intuitive—especially there are no self-intersections—after training with smoothed labels. Moreover, clusters ranked 1 and cluster ranked 5 are more separated compared with the cases utilizing regular OneHot labels. These observations help support the conclusion that the representations learned with the proposed smoothed labels are better able to model human intuition and logic on at least two aspects; i.e. **1**) representations corresponding to rankings from 1 through 5 are properly aligned on the representation manifold with their natural 1-D order, and **2**) clusters with rankings 1 and 5 are visually easier identified as the two ends on the representation manifold when compared to clusters with rankings 2 through 4, which comprise the interior points.

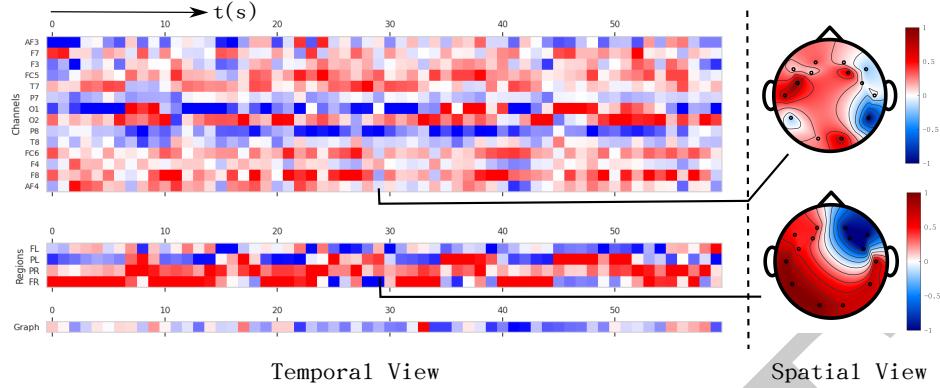


Figure 4: Deep features extracted in the temporal view and the spatial view at a time snapshot. From top to bottom: channel level, region Level and graph Level. The spatial view for the graph level is not shown since it is a scalar value.

3.2.2 Case Study – Deep Features

Fig. 4 provides a visual example of learned deep features before the classifier layer (i.e. a vector of 19 dimensions: 14[No. channels]+4[No. regions]+1[No. graphs] per 1s input as seen in Fig. 1), sequentially stacked along time representing the entire 1-minute recording (comprising 60 input samples) for subject S1 during the presentation of stimuli 12⁵. In the temporal view, feature values at different levels are normalized separately along the time dimension in the range [-1, 1]. The spatial view showcases features extracted at the snapshot at $t = 29 \sim 30$ seconds, which are further spatially re-normalized along all nodes to ensure uniformity. At the region level, nodes belonging to the same region are assigned with the same color as the region-level tomography plot and interpolated along regions, emphasizing larger scale/higher level spatial features.

Visualizations such as this can serve as a valuable tool to explore whether the learned patterns of the model align with clinical observations or real-world experiences. For instance, we analyze the spatial view by calculating the mean and standard deviation separately for male and female subjects, aggregating the results in Fig. 5. In this particular example, we observe some common patterns (e.g., high mean and standard deviation around F7 between Left Frontal/Temporal area, more complicated (pre)frontal pattern from female than male); however, there are distinct differences in feature patterns between male and female subjects. These observations appear to align with other numerical/clinical findings such as Peng et al. [2023], Hodgetts and Hausmann [2023].

An intriguing observation is that the mean pattern at the region level between the two sexes appears to be roughly inverted. Furthermore, when examining the standard deviation pattern at the region level, we observe that the color gradients in the male pattern tend to align along the anterior-posterior direction, whereas they align along the medial-lateral direction for the female. This may indicate a greater degree of asymmetry between the left and right brain hemispheres in females. While these numerical findings may not directly correspond to clinical experiences and require further investigation, they offer an interesting representation where the logical relation of “male-female” can be captured through simple arithmetic operations, such as taking the opposite for the mean pattern at the region level or rotating it by 90° for the standard deviation pattern. Additionally, if one possessed strong prior knowledge regarding the region patterns, the HiSTN design allows for its enforcement during training, leading to enhanced interpretability afterwards.

3.3 Subject-Independent Experiments

Classification tasks that are independent of the subject present more significant challenges compared to those that are dependent on the subject, primarily due to the introduction of additional complexities. For instance, discrepancies often arise when different individuals provide ratings in response to the same stimuli, with some extremes being a maximum rating of 5 reported by one person and a minimum rating of 1 reported by another - an example of which can be seen in Fig. 6. Moreover, the interpretation of the same score, such as 3, can vary between individuals based on their distinct personalities, signifying different emotional states. In experimental scenarios like Leave-One-Out Cross Validation (LOOCV), this inconsistency in labels, given identical input from different individuals, complicates the learning of effective features by the model.

⁵12 is chosen here for demonstration purposes only because it is one of the stimuli with the lowest standard deviation in terms of valence ratings provided from all subjects.

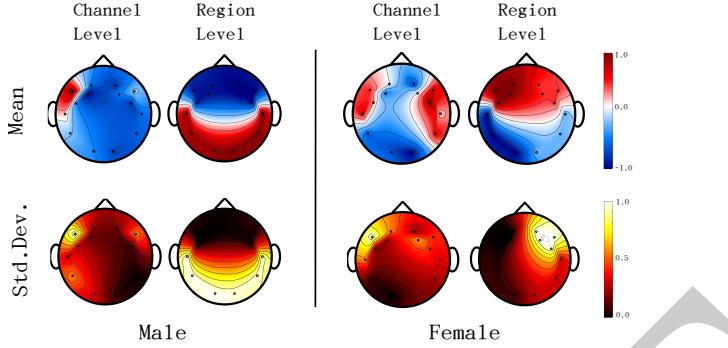


Figure 5: Channel level and region level patterns between male mean/std and female mean/std groups.

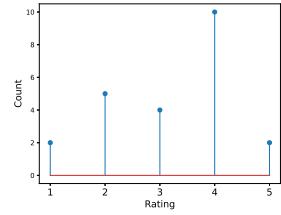


Figure 6: One example of distribution of self-reported valence ratings among all 23 subjects given the same stimuli (5th).

test data drawn from the remaining 50 seconds. During this second stage, the weights contained feature head, which mainly learn low-level features, are frozen. Only blocks deeper within the network are retrained. These include the network blocks after depthwise convolution for both EEGNet and the proposed HiSTN, and the dense layers after the final convolution blocks in each temporal and spatial branch for MSBAM. Following this process with each subject acting as a test case once, the model’s mean performance is computed and incorporated into Table 5.

For a more robust assessment, we continue to employ LOOCV for our experiments independent of subjects. In the case of the DREAMER dataset, the data from a single individual is chosen for testing each time, while data from the remaining subjects are combined for training purposes. While it is feasible to explore appropriate transfer learning strategies such as domain adaptation, these strategies constitute their own independent areas of research, encompass a wide range of topics and warrant considerable further investigation.

In this study, to align with our main objectives, we adopt a straightforward two-stage (pretraining/finetuning) framework to address the issue of label inconsistency across different subjects. In the *first* stage, during preprocessing, for the same trial where different subjects may report varying scores, we compute the prior score distribution among different subjects. We then adjust the label to the score with the highest likelihood and employ this corrected label for training. In the *second* stage, after the initial phase of training on other subjects’ data, we fine-tune the model using the initial 10-second data from the target subject. We then gather the performance on 1s-segment

Table 4: Hyper-parameters used in benchmark for subject-independent experiments.

Parameter	1st Stage	2nd Stage
Batchsize	120	100
LearningRate	0.01	0.001
Maximum Epochs	100	400

Subject-independent tasks, with their notably larger train/test gap, inherently present a more complex challenge, resulting in prediction performance that does not quite match the levels observed in subject-dependent experiments. Nonetheless, certain findings noted in these two tables align with earlier subject-dependent studies. Irrespective of the specific models utilized, the proposed special label smoothing consistently elevates the Seq2HR value significantly⁷. Moreover, under the same setting, Variation B (the regression task) proves more challenging to optimally train compared to the other variations. The proposed HiSTN-D configuration achieves a commendable equilibrium between pure accuracy metrics and prediction continuum, utilizing the fewest parameters. The performance decline from subject-dependent tasks to subject-independent tasks is more pronounced for MSBAM compared to the other two networks. This could be attributed to the constrained volume of data available for the second stage of training, making it harder to guide larger models such as MSBAM to adapt their prediction to specific test subjects.

⁷Similarly as seen in the subject dependent benchmark, pairwise t-test scores give very small p-values (i.e. EEGNet: $2.74e^{-16}/1.29e^{-12}$, MSBAM: $5.35e^{-13}/1.61e^{-9}$, HiSTN: $9.53e^{-14}/1.18e^{-11}$, for valence/arousal respectively, suggesting, e.g., strong statistical significance).

Table 5: Summary of subject-independent experiments on 5-classification tasks.

		Valence				Arousal			
Models	#Para.	F1(%)	Top2 Acc.(%)	Tri-P(%)	Seq2HR (%)	F1(%)	Top2 Acc.(%)	Tri-P(%)	Seq2HR (%)
EEGNet	A	3269	76.98 ± 5.91	91.95 ± 2.62	88.37 ± 4.29	35.68 ± 11.54	81.04 ± 6.98	93.53 ± 3.81	91.99 ± 6.13
	B	1217	50.26 ± 9.24	76.62 ± 7.56	86.44 ± 6.43	76.62 ± 7.56	57.88 ± 13.36	83.33 ± 11.13	92.05 ± 7.06
	C	8399	72.82 ± 8.01	88.25 ± 4.76	86.29 ± 5.11	48.67 ± 11.30	80.14 ± 7.36	93.22 ± 4.41	92.24 ± 6.93
	D	3269	76.51 ± 6.82	89.05 ± 3.80	89.65 ± 3.76	84.74 ± 4.89	77.94 ± 8.61	91.50 ± 6.39	92.91 ± 6.74
MSBAM	A	203,269	67.22 ± 7.25	84.85 ± 5.05	81.57 ± 5.21	36.63 ± 14.90	69.64 ± 7.79	88.08 ± 5.47	86.87 ± 7.56
	B	202,785	52.38 ± 7.22	74.49 ± 5.96	84.60 ± 4.45	74.49 ± 5.96	54.46 ± 10.42	80.02 ± 9.60	89.71 ± 6.98
	C	204,479	65.37 ± 9.34	83.54 ± 6.44	82.98 ± 4.44	52.75 ± 12.51	60.03 ± 14.26	83.01 ± 10.34	87.52 ± 7.65
	D	203,269	66.51 ± 7.11	82.12 ± 4.78	84.48 ± 4.44	75.53 ± 6.66	68.72 ± 8.54	85.81 ± 7.34	89.62 ± 6.93
HiSTN	A	1181	77.02 ± 5.63	91.23 ± 3.42	87.49 ± 4.04	37.18 ± 14.45	76.57 ± 8.54	92.18 ± 4.71	89.49 ± 7.95
	B	1101	55.69 ± 7.46	79.44 ± 6.58	89.09 ± 5.02	79.44 ± 6.58	62.63 ± 13.93	85.18 ± 10.80	93.08 ± 6.78
	C ⁶	1381	65.19 ± 6.21	84.91 ± 4.94	87.66 ± 5.00	70.87 ± 12.36	67.14 ± 8.32	88.15 ± 6.52	91.47 ± 6.95
	D	1181	78.34 ± 5.53	90.40 ± 3.02	90.59 ± 3.79	82.61 ± 5.99	81.59 ± 7.06	92.47 ± 5.18	93.61 ± 5.47

3.4 Ablation Study

3.4.1 Different choices of graphs

Fig 7 presents three distinct constructions of channel-level graphs, each of which corresponds to a regional level graph—specifically a 4-cycle, a 5-cycle, and a 3-cycle graph. In accordance with the HiSTN-D configuration, we implement an identical benchmark protocol for subject-independent studies as outlined in Section 3.3. The predictive performance for each construction is collated in Fig. 8. From our experimentation, the G_0 configuration emerged as the superior choice in terms of overall performance. Though numerical benchmarks can be close, different graph structures employed in

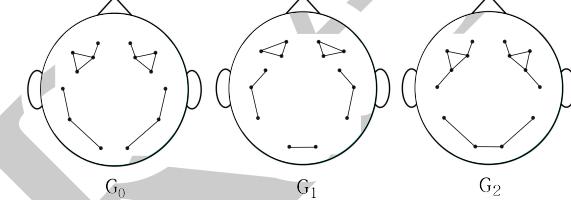


Figure 7: Different prior graph structures at the channel level.

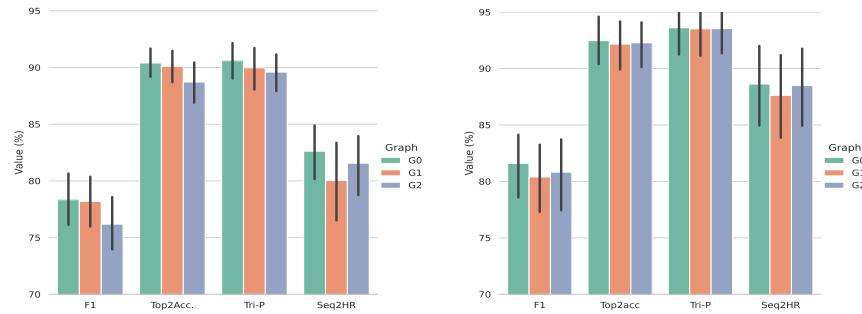


Figure 8: Performance when different prior graph structures are adopted. Left: Valence. Right: Arousal. Black lines represent the estimation of 95% confidence intervals.

HiSTN can yield varied patterns of deep features, an aspect lightly touched upon at the conclusion of Section 3.2. For a more tangible exploration, we use data from subject 12—chosen due to high prediction performance across all three variations—on stimuli 12. We illustrate the mean and standard deviation of the spatial view on the regional level in Fig. 9 (where both mean and standard deviation are calculated over time). As anticipated, different deep feature patterns emerge as a consequence of the varied choices in graph structures.

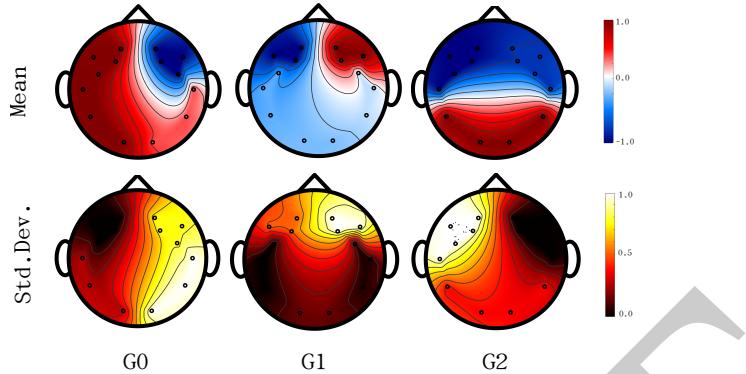


Figure 9: Patterns from the mean and standard deviation along the time direction for deep features learned at the region level when different graph structure is adopted in HiSTN.

3.5 Comparing with DiffPool

In Table 6, we evaluate the performance of HiSTN when the graph hierarchy (part II), as illustrated in Fig. 2, is replaced with DiffPool layers (designated as HiSTN(DP) in the table). This evaluation is carried out under the aforementioned subject-independent tasks, using the proposed label smoothing technique. The implementation of DiffPool layers were from Spektral Grattarola and Alippi [2021]. All other experimental parameters remained consistent with those outlined in Section 3.3. The table clearly demonstrates that a more specialized graph hierarchy design, as proposed in Section 2, facilitates improved predictions for this task.

Table 6: Comparison of mean test performance when graph hierarchy in HiSTN is replaced by DiffPool layers. The green numbers are p-values from the paired t-tests suggesting the improvement is of statistical significance.

Models	# Para.	Valence		Arousal	
		F1 (%)	Seq2HR(%)	F1 (%)	Seq2HR(%)
HiSTN(DP)-D	8848	67.22 ± 5.62	69.07 ± 11.92	68.59 ± 6.00	78.49 ± 11.13
HiSTN-D	1181	78.34 ± 5.53 ($\uparrow 1.07e^{-7}$)	82.61 ± 5.99 ($\uparrow 4.75e^{-7}$)	81.59 ± 7.06 ($\uparrow 7.35e^{-11}$)	88.62 ± 8.65 ($\uparrow 1.05e^{-8}$)

4 Discussion

The experiments conducted using the proposed HiSTN model have demonstrated the possibility for a lightweight yet thoughtfully constructed model to deliver effective prediction performance, even in the context of limited data. HiSTN’s hierarchical spatial and temporal architecture further allows for the integration of prior knowledge, thereby further help enhance human interpretability. This is particularly applicable in discerning the potentially meaningful spatial relations among recorded EEG data or in extracting different levels of information from brain function connectivity priors. Furthermore, our benchmark results indicate that when combined with other training techniques such as the proposed special label smoothing, HiSTN is capable of achieving better balance between quantitative and qualitative prediction. However, one must be mindful of the increased computational complexity brought on by the hierarchical design, especially during inter-layer message passing and node fusion, as compared to conventional convolution. Like other graph-based neural networks, it could be subject to common issues such as ‘over-squashing’ Alon and Yahav [2020] or ‘over-smoothing’ Li et al. [2018].

This research sets a stage for further exploration in numerous fields with the potential for significant advancements in application performance. Notably, the employed graphs and hierarchical designs could be refined through a blend of clinical knowledge and mathematical tools. Theories on extending concepts from smooth manifolds such as Ricci flows to graph-like discrete structures could potentially help optimize a graph structure initially created based on clinical priors for deep learning purposes Topping et al. [2021]. Concurrently, the strategic implementation of efficient transfer learning techniques might enable a seamless translation of knowledge acquired from the training domain to specific target subjects. Additionally, the challenge of label inconsistency, which can be seen as a noisy label issue or fuzzy label issue, could be more effectively tackled using reinforcement/contrastive learning techniques or be examined under fuzzy set/logic framework. These approaches, nested within the semi-supervised learning framework, have the potential to better manage the problem by appropriately weighting or selectively choosing information-rich samples.

5 Conclusion

This paper showcases through both subject-dependent and independent experiments on the DREAMER under the finer-grained 5-classification tasks that the proposed HiSTN can offer a highly parameter-efficient solution. By integrating the proposed spatial label smoothing technique, the quality of the model's predictions can be significantly enhanced, as indicated by the high likelihood of top predictions encapsulating true labels and their proximity to each other. Despite certain limitations and potential future areas for exploration highlighted in the Discussion Section, this study serves as a promising step towards optimizing the balance between quantitative metrics and qualitative behavior in model predictions, particularly in scenarios where data is scarce and parameter efficiency is a critical factor.

S.1: Generally message passing does not commute with temporal convolutions

Given the feature matrix $X \in \mathcal{R}^{C \times T}$ generated by stacking signals of length T from C different channels, we let A denote the matrix multiplied from the left for message passing (ψ), and $W \in \mathcal{R}^{C \times k}$ be the stack of convolution kernels of length k . The channel-wise time convolution $*_t$ defined for 2d feature matrix is then performed as follows:

$$\mathcal{T}(X) = X *_t W = \begin{bmatrix} X_{1\cdot} * W_{1\cdot} \\ \vdots \\ X_{C\cdot} * W_{C\cdot} \end{bmatrix}$$

where $*$ is the regular 1d convolution along time and entries $[]_i$. (e.g. $X_{1\cdot}$) represent the entirety of the i th row vector. We can then calculate $\mathcal{T} \circ \psi = (AX) *_t W$ and $\psi \circ \mathcal{T} = A(X *_t W)$ as follows:

$$\mathcal{T} \circ \psi(X) = (AX) *_t W = \begin{bmatrix} \sum_{j=1}^C A_{1j} X_{j\cdot} \\ \vdots \\ \sum_{j=1}^C A_{Cj} X_{j\cdot} \end{bmatrix} *_t W \quad (7)$$

$$= \begin{bmatrix} \sum_{j=1}^C A_{1j} X_{j\cdot} * W_{1\cdot} \\ \vdots \\ \sum_{j=1}^C A_{Cj} X_{j\cdot} * W_{C\cdot} \end{bmatrix}, \quad (8)$$

$$\psi \circ \mathcal{T}(X) = A(X *_t W) = A \begin{bmatrix} X_{1\cdot} * W_{1\cdot} \\ \vdots \\ X_{C\cdot} * W_{C\cdot} \end{bmatrix} \quad (9)$$

$$= \begin{bmatrix} \sum_{j=1}^C A_{1j} X_{j\cdot} * W_{j\cdot} \\ \vdots \\ \sum_{j=1}^C A_{Cj} X_{j\cdot} * W_{j\cdot} \end{bmatrix}. \quad (10)$$

Thus, in order to have $(AX) *_t W = A(X *_t W)$, one must require that,

$$\sum_{j=1}^C A_{ij} X_{j\cdot} * W_{i\cdot} = \sum_{j=1}^C A_{ij} X_{j\cdot} * W_{j\cdot}, \quad \forall i. \quad (11)$$

or equivalently,

$$\sum_{j=1}^C A_{ij} X_{j\cdot} * (W_{i\cdot} - W_{j\cdot}) = 0, \quad \forall i. \quad (12)$$

The following example gives a straightforward calculation.⁸ Consider the matrices:

$$A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, X = \begin{bmatrix} 1 & 3 & -1 & -2 \\ -1 & 2 & 1 & 0 \end{bmatrix}, W = \begin{bmatrix} -1 & 2 \\ 3 & 1 \end{bmatrix},$$

so that,

$$(AX) *_t W = \begin{bmatrix} 0.5 & 4 & -0.5 & -2 \\ -0.5 & 3.5 & 0.5 & -1 \end{bmatrix} *_t \begin{bmatrix} -1 & 2 \\ 3 & 1 \end{bmatrix} \quad (13)$$

$$= \begin{bmatrix} 7.5 & -5 & -3.5 \\ 2 & 11 & 0.5 \end{bmatrix} \quad (14)$$

while,

$$A(X *_t W) = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \begin{bmatrix} 5 & -5 & -3 \\ -1 & 7 & 3 \end{bmatrix} \quad (15)$$

$$= \begin{bmatrix} 4.5 & -1.5 & -1.5 \\ 1.5 & 4.5 & 1.5 \end{bmatrix}, \quad (16)$$

thus arriving with $(AX) *_t W \neq A(X *_t W)$.

References

- R. Adolphs. Neural systems for recognizing emotion. *Current opinion in neurobiology*, 12(2):169–177, 2002.
- U. Alon and E. Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2020.
- Aristotle. *Rhetoric*. Modern Library, New York, 1984. ISBN 978-0394604572.
- H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, and X. Chen. EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowledge-Based Systems*, 205:106243, 2020.
- M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.
- P. Ekman. Darwin’s contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3449–3451, 2009.
- Epictetus. *Enchiridion*. Dover Publications, 2004. ISBN 978-0486433592.
- C. A. Gabert-Quillen, E. E. Bartolini, B. T. Abravanel, and C. A. Sanislow. Ratings for emotion film clips. *Behavior research methods*, 47:773–787, 2015.
- A. Ghosh, H. Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. *Proceedings of the AAAI conference on artificial intelligence*, 31(1), 2017.
- S. Gong, K. Xing, A. Cichocki, and J. Li. Deep learning in EEG: Advance of the last ten-year critical period. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):348–365, 2021.
- D. Grattarola and C. Alippi. Graph neural networks in tensorflow and keras with spektral [Application Notes]. *IEEE Computational Intelligence Magazine*, 16(1):99–106, 2021. doi:10.1109/MCI.2020.3039072.
- C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/guo17a.html>.
- M. He, Z. Wei, and J.-R. Wen. Convolutional neural networks on graphs with chebyshev approximation, revisited. *arXiv preprint arXiv:2202.03580*, 2022.

⁸The term *convolution* referred to in the neural network setting is actually a *correlation* in standard mathematics terminology, i.e. the kernel is not rotated by 180°. In the example, we followed the neural network setting, but one can easily verify that the equation does not hold for either *correlation* or *convolution*.

- S. Hodgetts and M. Hausmann. *Sex/Gender Differences in Brain Lateralisation and Connectivity*, pages 71–99. Springer International Publishing, Cham, 2023. ISBN 978-3-031-26723-9. doi:10.1007/978-3-031-26723-9_3. URL https://doi.org/10.1007/978-3-031-26723-9_3.
- X. Ju, M. Li, W. Tian, and D. Hu. EEG-based emotion recognition using a temporal-difference minimizing neural network. *Cognitive Neurodynamics*, pages 1–12, 2023.
- S. Katsigiannis and N. Ramzan. Dreamer: A database for emotion recognition through EEG and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, 22(1):98–107, 2017.
- 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*, 3(1):18–31, 2011.
- D. Kuang, C. Michoski, W. Li, and R. Guo. From gram to attention matrices: a monotonicity constrained method for EEG-based emotion classification. *Applied Intelligence*, pages 1–20, 2023.
- V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.
- J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam. Emotion and decision making. *Annual review of psychology*, 66: 799–823, 2015.
- C. Li, B. Wang, S. Zhang, Y. Liu, R. Song, J. Cheng, and X. Chen. Emotion recognition from EEG based on multi-task learning with capsule network and attention mechanism. *Computers in Biology and Medicine*, 143:105303, 2022a.
- D. Li, L. Xie, B. Chai, Z. Wang, and H. Yang. Spatial-frequency convolutional self-attention network for EEG emotion recognition. *Applied Soft Computing*, 122:108740, 2022b.
- Q. Li, Z. Han, and X.-M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI conference on artificial intelligence*, 32(1), 2018.
- Q. Li, T. Zhang, C. P. Chen, K. Yi, and L. Chen. Residual GCB-Net: Residual graph convolutional broad network on emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2022c.
- X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen. Eeg based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4):1–57, 2022d.
- Y. Liu, Y. Wei, C. Li, J. Cheng, R. Song, and X. Chen. Bi-CapsNet: A binary capsule network for EEG-based emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 2022.
- D. Maheshwari, S. K. Ghosh, R. Tripathy, M. Sharma, and U. R. Acharya. Automated accurate emotion recognition system using rhythm-specific deep convolutional neural network technique with multi-channel eeg signals. *Computers in Biology and Medicine*, 134:104428, 2021.
- R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.
- D. Peng, W.-L. Zheng, L. Liu, W.-B. Jiang, Z. Li, Y. Lu, and B.-L. Lu. Identifying sex differences in eeg-based emotion recognition using graph convolutional network with attention mechanism. *Journal of Neural Engineering*, 20(6):066010, 2023.
- R. W. Picard. *Affective computing*. MIT press, 2000.
- R. Plutchik. *Emotions and life: Perspectives from psychology, biology, and evolution*. American Psychological Association, 2003.
- D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Affect recognition from scalp-EEG using channel-wise encoder networks coupled with geometric deep learning and multi-channel feature fusion. *Knowledge-Based Systems*, 250:109038, 2022.
- J. Quan, Y. Li, L. Wang, R. He, S. Yang, and L. Guo. EEG-based cross-subject emotion recognition using multi-source domain transfer learning. *Biomedical Signal Processing and Control*, 84:104741, 2023.
- J. A. Russell. Affective space is bipolar. *Journal of personality and social psychology*, 37(3):345, 1979.
- T. Sainburg, L. McInnes, and T. Q. Gentner. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907, 2021.
- Seneca. *On Anger: De Ira*. Independently published, 2017. ISBN 978-1521777268.
- M. K. Singh, M. Singh, et al. A deep learning approach for subject-dependent & subject-independent emotion recognition using brain signals with dimensional emotion model. *Biomedical Signal Processing and Control*, 84:104928, 2023.

- H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- T. Song, W. Zheng, P. Song, and Z. Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.
- T. Song, W. Zheng, S. Liu, Y. Zong, Z. Cui, and Y. Li. Graph-embedded convolutional neural network for image-based EEG emotion recognition. *IEEE Transactions on Emerging Topics in Computing*, 10(3):1399–1413, 2021.
- R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieri, S. Winkler, and N. Sebe. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2016.
- C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing*, 2020.
- W. Tian, M. Li, X. Ju, and Y. Liu. Applying multiple functional connectivity features in GCN for EEG-Based human identification. *Brain Sciences*, 12(8):1072, 2022.
- J. Topping, F. Di Giovanni, B. P. Chamberlain, X. Dong, and M. M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. *arXiv preprint arXiv:2111.14522*, 2021.
- M. Wang, H. El-Fiqi, J. Hu, and H. A. Abbass. Convolutional neural networks using dynamic functional connectivity for EEG-based person identification in diverse human states. *IEEE Transactions on Information Forensics and Security*, 14(12):3259–3272, 2019.
- Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 2022.
- Z. Wang, T. Gu, Y. Zhu, D. Li, H. Yang, and W. Du. FLDNet: Frame-level distilling neural network for EEG emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2533–2544, 2021.
- Y. Wei, Y. Liu, C. Li, J. Cheng, R. Song, and X. Chen. Tc-net: A transformer capsule network for EEG-based emotion recognition. *Computers in Biology and Medicine*, 152:106463, 2023.
- Y. Wu, M. Xia, L. Nie, Y. Zhang, and A. Fan. Simultaneously exploring multi-scale and asymmetric EEG features for emotion recognition. *Computers in Biology and Medicine*, 149:106002, 2022.
- Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31, 2018.
- 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*, 2021.
- 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*, 7(3):162–175, 2015.

DECAN: A Denoising Encoder via Contrastive Alignment Network for Dry Electrode EEG Emotion Recognition

Meihong Zhang, Shaokai Zhao, Shuai Wang, Zhiguo Luo, Liang Xie, Tiejun Liu, Dezhong Yao, *Senior Member, IEEE*, Ye Yan and Erwei Yin

Abstract—EEG signal is important for brain-computer interfaces (BCI). Nevertheless, existing dry and wet electrodes are difficult to balance between high signal-to-noise ratio and portability in EEG recording, which limits the practical use of BCI. In this study, we propose a Denoising Encoder via Contrastive Alignment Network (DECAN) for dry electrode EEG, under the assumption of the EEG representation consistency between wet and dry electrodes during the same task. Specifically, DECAN employs two parameter-sharing deep neural networks to extract task-relevant representations of dry and wet electrode signals, and then integrates a representation-consistent contrastive loss to minimize the distance between representations from the same timestamp and category but different devices. To assess the feasibility of our approach, we construct an emotion dataset consisting of paired dry and wet electrode EEG signals from 16 subjects with 5 emotions, named PaDWEED. Results on PaDWEED show that DECAN achieves an average accuracy increase of 6.94% comparing to state-of-the art performance in emotion recognition of dry electrodes. Ablation studies demonstrate a decrease in inter-class aliasing along with noteworthy accuracy enhancements in the delta and beta frequency bands. Moreover, an inter-subject feature alignment can obtain an accuracy improvement of 5.99% and 5.14% in intra- and inter-dataset scenarios, respectively. Our proposed method may open up new avenues for BCI with dry electrodes. PaDWEED dataset used in this study is freely available at <https://huggingface.co/datasets/peiyu999/PaDWEED>.

Index Terms—Emotion recognition, EEG, dry electrode, contrastive learning.

I. INTRODUCTION

EMOTION plays a crucial role in human behavior, decision-making, social interactions, and overall well-

This work was supported in part by the grants from the National Natural Science Foundation of China under Grant 62332019, 62076250 and 62406338, the National Key Research and Development Program of China (2023YFF1203900, 2023YFF1203903). (Corresponding authors: Shaokai Zhao; Erwei Yin.)

Meihong Zhang, Shuai Wang are with School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China. And Defense Innovation Institute, Academy of Military Sciences (AMS), Beijing, China. (e-mail: zmh@std.uestc.edu.cn; tjwangshuai1234@163.com).

Shaokai Zhao, Zhiguo Luo, Liang Xie, Ye Yan, Erwei Yin are with Defense Innovation Institute, Academy of Military Sciences (AMS), Beijing, China. (e-mail: lnkzsk@yeah.net; 963619079@qq.com; xielnudt@gmail.com; yynudt@126.com; yinerwei1985@gmail.com).

Tiejun Liu, Dezhong Yao are with The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China. And School of Life Science and Technology, Center for Information in Medicine, University of Electronic Science and Technology of China, Chengdu, China. (e-mail: liutiejun@uestc.edu.cn; dyao@uestc.edu.cn).

being. In recent years, emotion recognition has sparked significant interdisciplinary interest from fields ranging from psychology to engineering [1], based on multiple physiological signals [2]–[5]. The cause appears to be, in part, that the signals mentioned above are difficult to disguise, and hence they reflect genuine emotional states [6]. Electroencephalogram (EEG) is considered as an effective and reliable neural signal that carry the information about different emotions, as it objectively records our brain activity, which serves as the central nervous system for emotion processing [7].

EEG signals are extremely weak, which are easily susceptible to environmental noise and other electrophysiological signals [8]. Wet electrodes, regarded as the gold standard for EEG devices and clinical EEG recordings [9], [10], offer the relative higher signal-to-noise ratio, have yielded numerous noteworthy research findings in the field of brain-computer interface [11]–[13]. Although they have achieved many noteworthy research results in the field of brain-computer interface, the complex preparation and cleaning procedures have severely limited the practical application of BCI in real-world scenarios, prompting exploration into dry electrode systems [14]–[16].

However, the low signal-to-noise ratio of dry EEG systems presents a challenge for portable emotion recognition applications. Inspired by the achievements by wet electrode EEG systems, we propose to enhance the emotion recognition accuracy of dry EEG recordings by leveraging insights gained from wet EEG data. Our novel approach involves training a Denoising Encoder using a Contrastive Alignment Network (DECAN) to extract task-specific information embedded within dry electrode signals with the assistance of wet electrodes which is not necessary in test phase. The main contributions of this study can be summarized as follows:

- We propose a Denoising Encoder by Contrastive Alignment Network (DECAN) to enhance the recognition performance of dry electrode signals by leveraging the knowledge learned from wet electrode data, which is comprised of two partially-shared deep neural networks (DNN) for efficient emotional feature extraction and a feature alignment contrastive learning strategy.
- We construct a new dataset consists of Paired Dry and Wet Electrode EEG Data (PaDWEED) collected from 16 subjects under identical video stimuli and experimental conditions to validate the effectiveness of our DECAN in enhancing dry-electrode performance. Additionally, we provide baseline performance metrics for PaDWEED

which can also be set as a benchmark for the dry electrode EEG emotion recognition task.

- Experiments on PaDWEEED show that our proposed DECAN achieves state-of-the-art results on dry electrode EEG emotion recognition, and inter-subject improvements observed in both the intra- and inter-dataset feature alignment tasks demonstrate its ability to overcome more challenging scenarios.

The layout of the paper is as follows. Section 2 provides a review of previous research on emotion recognition using dry electrodes. Additionally, a concise overview of existing EEG emotion recognition databases is also presented. In Section 3, a detailed explanation of the materials and protocols used for our dataset constructing is provided. Section 4 and Section 5 introduce our proposed DECAN model and extensive experiment results. Section 6 and 7 are dedicated to the discussion of the findings and the conclusion of the study respectively.

II. RELATED WORKS

A. EEG emotion recognition based on dry electrode

The portability of dry electrode EEG system enables the potential for brain-computer emotion recognition applications in daily life. However, its low signal-to-noise ratio results in poor system performance. Currently, two mainstream solutions have been proposed to mitigate this issue. From a hardware perspective, enhancements are made to the dry electrode EEG acquisition system. This includes optimizing the electrode structure and replacing the electrode material to reduce contact impedance, thereby enhancing the signal-to-noise ratio of the acquired signal. So far, there are several mature commercial dry electrode devices available for EEG data acquisition in emotion recognition, such as the DSI-24, OpenBCI, EMOTIV EPOC and so on. In another study, a four-channel textile cap was designed with dry electrodes secured by an ultra-soft gel holder while introducing stylish and ergonomic design features to enhance wearability and comfort. The average binary emotion classification accuracy was found to be 81.32% among five healthy elderly participants [17].

From a software perspective, the performance of the dry electrode EEG system has been enhanced by optimizing the recognition algorithm, including feature extraction, classifier design. Lakhan et al. recruited 43 subjects to watch pre-labeled emotional visual stimuli, while using OpenBCI and Empatica4 to collect EEG and peripheral physiological signals. Through a classification algorithm based on K-means clustering, they achieved accuracy rates of 70% for arousal and 67% for valence [18]. Katsigiannis et al. simultaneously collected EEG and Electrocardiogram (ECG) data from subjects while they were viewing movie clips. When using unimodal features, the SVM-RBF classification method achieved higher accuracy rates for valence, arousal, and dominance using EEG features compared to ECG features. The accuracy rates for valence, arousal, and dominance were 62.49%, 62.17%, and 61.84% respectively [19]. Javaid et al. employed the SVM-RBF algorithm to classify EEG data collected using OpenBCI, obtained an accuracy of 87.62% for arousal and 83.28% for valence [20]. Lan et al. collected EEG signals

from individuals with major depressive disorder and healthy control subjects. They utilized the topological information among EEG channels for emotion recognition and depression detection. The study evaluated the promising ability of the emotional EEG patterns in distinguishing individuals with depression from the healthy control group [16]. Xu et al. put forward a novel framework for emotion recognition in VR emotional scenes using EEG signals. They employed feature extraction techniques in the time domain, frequency domain, and spatial domain from the EEG data. The extracted features were then utilized to train an ensemble model using the Model Stacking approach, combining gradient boosting decision tree, random forest, and SVM models. The resulting average accuracy achieved for classifying positive and negative emotions was approximately 81.30% [21]. However, given that research on dry electrode emotion recognition is still in its nascent stages, the methods adopted in the aforementioned studies primarily rely on traditional signal processing and machine learning algorithms. The extracted EEG features are relatively shallow, and primarily focusing on the recognition of emotion dimension models. In contrast, this study focuses on dry electrode EEG discrete emotion recognition, employing deep learning methods to leverage wet electrode signals with relatively higher signal-to-noise ratio to enhance the performance of dry electrode EEG emotion recognition.

B. Available databases for EEG-based emotion recognition

There has been a consecutive release of emotion databases that containing EEG, within its range this decade, which each of them serves to distinctive experimental motivation [22]. Specific information of these databases has been reported in Table I. As an early released multimodal public dataset containing both physiological responses and facial expressions, MAHNOB-HCI has garnered significant attention [23], and numerous studies have been conducted to validate algorithm performance using this dataset [24]–[26]. The DEAP dataset, released during the same period, explores the possibility of classifying emotions induced by music videos, where this type of stimulus has never been explored before [27]. Also from the perspective of selecting stimulus materials, Song et al. eliminated the impact of cultural dependence on the induction of discrete emotions through large-scale and rigorous screening of materials. Finally, they elaborately selected 28 videos as standardized elicitation samples and recorded the physiological responses of subjects while watching the above videos. signal, thereby constructing the MPED database [28].

Changes in participants' physiological states over time inevitably lead to a decline in emotion classification performance, posing challenges for practical applications. To address this issue, Zheng et al. introduced a novel dataset called SEED and aimed to discover stable patterns within repeated sessions of the same participant. The results obtained indicated that neural patterns during training and across different training instances were relatively stable [29]. Taking a similar perspective on the influence of time on emotions, Sharma et al. proposed and developed the CASE database, which comprises continuous annotations of emotions and data from multiple

TABLE I
PUBLICLY AVAILABLE DATABASES FOR EEG-BASED EMOTION RECOGNITION

Database	Stimuli	Participants	Emotion states	EEG Electrode	Motivation
MAHNOB-HCI	20 video clips	27	Valence, Arousal, Dominance, Predictability, Anger, Anxiety, Fear, Sadness, Disgust, Neutrality, Surprise, Amusement, Joy	Wet	Multimodal synchronization database.
DEAP	40 music video clips	32	Valence, Arousal, Dominance, Liking, Familiarity	Wet	Using music video as stimulus for the first time.
ASCERTAIN	36 movie clips	15	Negativity, Neutrality, Positivity	Wet	Exploring the influence of individual traits on emotion classification using multimodal physiological signals.
MPED	28 video clips	23	Anger, Fear, Sadness, Disgust, Neutrality, Funny, Joy	Wet	Eliminating the impact of cultural background differences on multimodal databases through large-scale rigorous screening.
SEED	15 film clips	15	Negativity, Neutrality, Positivity	Wet	Focusing on the stability of multimodal emotional physiological signals.
CASE	8 film clips	30	Valence, Arousal; Amusing, Boring, Relaxing and Scary	Wet	Real-time annotation.
DREAMER	18 video clips	23	Valence, Arousal, Dominance	Dry	First database containing EEG and ECG signal recordings from low-cost, off-the-shelf, portable wireless devices.
PaDWEED (Ours)	25 film clips	16	Anger, Fear, Sadness, Happiness, Neutrality	Dry, Wet	Multimodal database for discrete emotions, using both dry and wet electrodes to collect EEG signals for the first time.

physiological sensors [30]. They posited that emotions are dynamic phenomena that evolve over time in response to stimuli [31], thus emphasizing the examination of the temporal nature of emotional changes.

Taking into consideration that emotions are highly subjective phenomena and are influenced by various factors including personality, background, and psychological factors, Subramanian *et al.* proposed the ASCERTAIN database, which is the first database to link personality traits and emotional states through physiological responses. The database includes big-five personality scales, self-ratings of emotions from 58 users, as well as their physiological and facial activity data [32]. In addition, considering the integration of affective computing with various daily applications, the DREAMER dataset has been proposed, which is a database consisting of EEG and ECG signal records collected by portable devices, aiming at identifying the affective state after each stimuli, in terms of valence, arousal, and dominance [19].

In this paper, to better support our research, we develop a new EEG dataset which includes Paired Dry and Wet electrode EEG Data (PaDWEED). This dataset is obtained from 16 participants who watch the same film clip stimuli in two separate experiments. In addition to EEG, physiological measurements such as ECG, electrooculogram (EOG), blood volume pulse (BVP), galvanic skin response (GSR), respiration (RSP), and skin temperature (SKT) are also recorded. To the best of our knowledge, this dataset represents a pioneering

effort by offering both wet and dry electrode EEG signals collected from the same set of participants, which is valuable for understanding the underlying mechanisms and developing classification algorithms for emotion recognition using dry electrodes.

III. DATASET CONSTRUCTION

We build a new EEG emotion dataset with Paired Dry and Wet Electrode EEG Data (PaDWEED) which differs from existing publicly available datasets, to support our research. In our experiment, we recruit the same subjects to participant two separate sessions of both dry and wet electrode system experiments. For each subject, the order of the experiments is not predetermined and there is a minimum interval of two weeks between the two sessions, or a longer duration. A summary of the compiled data is provided in Table II.

A. Ethics statement

All subjects participated in the experiment on the basis of understanding the experimental procedures and equipment safety, and signed an informed consent form before the experiment. The experiment is conducted in accordance with the guidelines of the Declaration of Helsinki and is approved by the ethics committee of the University of Electronic Science and Technology of China.

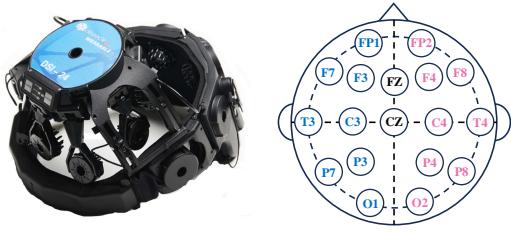


Fig. 1. The DSI-24 EEG cap and the sensor layout with 18 channels.

B. Experiment Setup

Stimulus selection. In our emotion experiments, we select predefined Chinese movie clips as stimuli for emotion elicitation [33]. This choice is made considering the potential influence of native cultural factors on emotion elicitation in experimental settings [34], [35]. Specifically, the 25 videos correspond to the emotions of anger, fear, sadness, happiness, neutrality, with an equal number of stimuli in each emotion category. The length of each edited clip is approximately 2.5 minutes. The selection of these video materials is based on three criteria: (a) stability of video content, (b) retention of characters in the scenes (excluding neutral materials), and (c) the absence of simultaneous presence of positive and negative emotions. Each video clip is edited to ensure consistency between the scenes and emotional content throughout the entire performance.

Materials. Two PCs are used, one for stimulus presentation, positioned approximately one meter in front of the user, and another for recording data, allowing the experimenter to verify the recorded sensor data. In the wet electrode system experiment, an ESI NeuroScan System2 with a 62-channel active AgCl electrode cap is used to collect EEG data with a sampling frequency of 1000 Hz. In the dry electrode experiment, we utilize the Dry Sensor Interface (DSI-24) and DSI-Streamer to record EEG signals at a sampling rate of 300 Hz. The sensors are positioned according to the international 10-20 system, with the default setting of the Pz electrode as the reference. The DSI-24 EEG cap and sensor layout for 18 channels are illustrated as Fig. 1.

Although peripheral physiological signals are not the focus of this study, we still utilize the MP160 data acquisition system to collect ECG, GSR, PPG, RSP, and ST signals at a sampling rate of 1000 Hz, which have demonstrated promising performance in emotion estimation research, therefore may be useful for other studies. The stimulus presentation protocol is developed using MATLAB's Psychtoolbox. Synchronization

markers are sent from the stimulus presentation PC to the physiological data recorder to mark the start and end of each stimulus.

Participants. A total of 16 volunteers (10 males, 6 females) with an average age of 23.5 ± 1.71 years participate in both data collection experiments. All participants are recruited from the Tianjin (Binhai) Artificial Intelligence Civil-Military Integration Innovation Center and are university students. They self-reported normal vision or corrected-to-normal vision and normal hearing. Prior to the experiment, participants are informed about the experimental procedure and instructed to sit comfortably and attentively watch the upcoming movie clips without diverting their attention from the screen and to minimize any noticeable movements.

Protocol. Each participant performs the experiment in a session lasting approximately 75 minutes in which we define each viewing of a video clip as one trial. The experiment begins with a three-minute baseline recording during which participants are shown a fixed cross. The video clips are then divided into 5 blocks, with each block consisting of 5 trials. Each trial included a 30-second introductory prompt before each clip and a 30-second rest period after each clip. During both the introductory prompt and the rest period, participants are given the freedom to decide whether to wait or proceed to the next stage based on their own condition. Fig. 2 presents the detailed protocol.

IV. METHODOLOGY

A. Problem setup.

In dry electrode EEG emotion recognition systems, poor recognition performance can be attributed to lower signal-to-noise ratio, as weaker and noisier signals may hinder the detection of subtle emotion patterns and affect the overall recognition performance. Suppose $X_w = \{(x_{wi}^j, y_{wi}^j), \forall i \in [1, r], \forall j \in [1, s]\}$ and $X_d = \{(x_{di}^j, y_{di}^j), \forall i \in [1, r], \forall j \in [1, s]\}$ denote the set of input data and labels of wet and dry electrode signals, where i and j denote the sample and participant indices respectively. r is the number of samples belonging to each participant and s is the total number of participants. Based on previous observations indicating the presence of similar valuable information in both wet electrode and dry electrode systems, our objective is to enhance the accuracy of dry electrode EEG emotion recognition by leveraging the valuable information obtained from wet electrode EEG signals.

B. Solution overview

As illustrated in Fig. 3, our proposed method is a contrastive learning-based architecture. Specifically, we utilize DNN models that have shown superior performance in previous analysis to learn high-level features from both wet and dry electrode EEG signals. Subsequently, we employ contrastive learning to perform pairwise representation alignment (wet versus dry). In the following subsections, we provide a detailed description of each step in our proposed method.

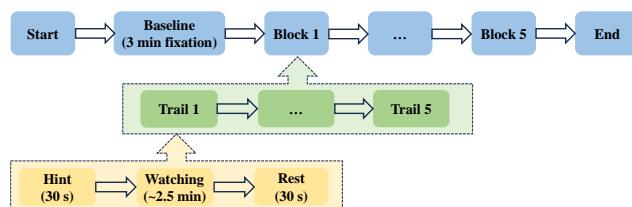


Fig. 2. The protocol used in both wet and dry electrode emotion experiments.

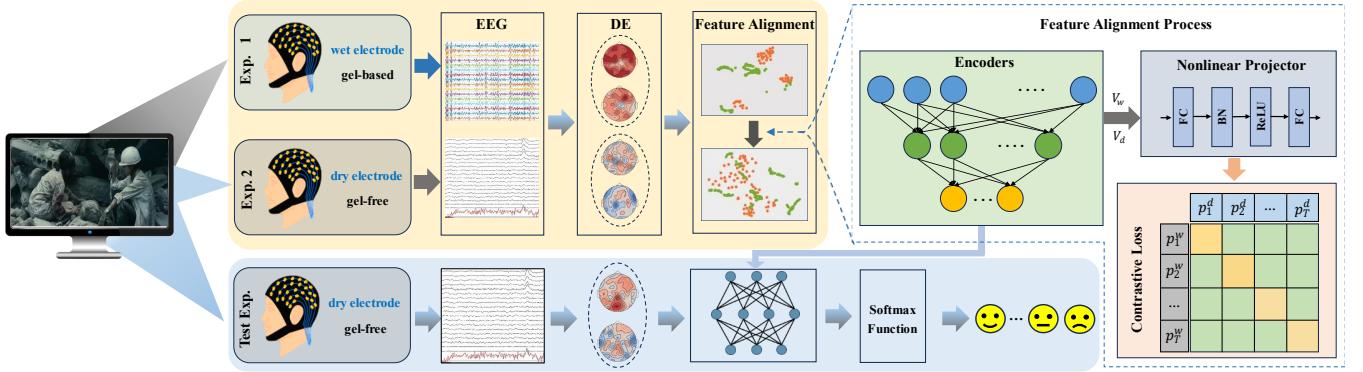


Fig. 3. The architecture of our proposed framework DECAN with pairwise representation alignment with contrastive learning for emotion recognition.

C. Our approach

Feature extraction. Considering the impact of the length of samples on recognition performance [12], [36], we divide a trial into five-second non-overlapping segments to ensure a balanced number and sample length of training samples, where each segment is regarded as a model training sample. The recorded EEG signals are susceptible to contamination from noise and artifacts but not of cerebral origin. Therefore, a series of preprocessing steps are adopted in our research to improve the EEG quality. First, a bandpass filter from 1 Hz to 50 Hz is applied to the raw EEG signals acquired from both dry and wet electrodes. This filtering approach aims to preserve the frequency range of the EEG signals containing emotional information while simultaneously removing any DC offsets and high-frequency interference such as muscle activity or other artifacts. Next, a notch filter at 50 Hz is employed to eliminate the power line interference caused by AC electrical sources or poor grounding connections. Finally, the filtered signals are downsampled to 200 Hz to reduce the complexity of subsequent signal processing and accelerate the overall signal processing speed.

It is well documented that the differential entropy (DE) is of efficient EEG feature for human emotional states [37], [38]. Hence, it is employed in this research to facilitate the subsequent classification model. Specifically, the preprocessed EEG signals are separated into five frequency band: delta (1-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-31 Hz) and gamma (31-50 Hz), then the DE for Gaussian distribution can be extracted as follows:

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

where X denotes the Gaussian distribution $N(\mu, \sigma^2)$, π and e are constants.

Assuming emotional states are situated in a continuous space and gradual transition, the linear dynamical system is a popular feature smoothing technique widely employed to filter out components irrelevant to emotional states [38]–[40]. Given the impressive performance demonstrated by this method in various studies, we also employ it to smooth the extracted DE features.

Encoders. Given a series of paired wet and dry electrode EEG signals X_w and X_d collected from same participant during the same stimulus segment, we use the extracted DE features from multiple frequency bands $F_w = \{f_{iw}^j, \forall i \in [1, r], \forall j \in [1, s]\}$ and $F_d = \{f_{id}^j, \forall i \in [1, r], \forall j \in [1, s]\}$ as inputs to the encoder and output high-level features $V_w = \{v_{iw}^j, \forall i \in [1, r], \forall j \in [1, s]\}$, $V_d = \{v_{id}^j, \forall i \in [1, r], \forall j \in [1, s]\}$. Specifically, F_w and F_d are extracted from the raw signal after filtering and downsampling, as described previously. Then we choose to utilize two partial weight-sharing DNN as the encoders in our approach. This particular architecture has consistently shown superior performance compared to other models in extracting paired features effectively. To promote efficient feature extraction, we have implemented weight sharing by sharing the last two linear sublayers and the projector between the two types of signals.

$$V_w = DNN_w(F_w) \quad (2)$$

$$V_d = DNN_d(F_d) \quad (3)$$

Feature alignment. Contrastive learning is a highly effective method for self-supervised representation learning, demonstrating significant achievements in pairwise feature learning across diverse domains [41]–[43]. Recent advancements in neuroscience have provided inspiration, suggesting that contrastive learning can be applied to extract subtle information from signals within the central nervous system [12]. This growing body of evidence highlights the potential of utilizing contrastive learning in neural research. Based on the aforementioned rationale, we have chosen to employ contrastive learning to accomplish the task of aligning features between wet and dry electrode EEG signals, details are introduced as follows.

Projectors. Motivated by the SimCLR framework, which suggests that nonlinear projection of encoded features can yield better performance in downstream tasks compared to linear projection or no operation [43], we have opted to incorporate a structure comprised of fully connected layers

and a ReLU layer. This configuration allows us to obtain the desired projection features, shown in Fig.3.

$$P_w = \text{Projector}(V_w) \quad (4)$$

$$P_d = \text{Projector}(V_d) \quad (5)$$

The contrastive loss. Given a paired of sequences of projection features $P_w = p_i^w, p_{i+1}^w, \dots, p_T^w$ and $P_d = p_i^d, p_{i+1}^d, \dots, p_T^d$, where T is the number of samples, w and d denote wet and dry electrode respectively. We define samples as positive when they originate from the same subject and different EEG collection systems, while also corresponding to the same time segment of the same emotion type. Conversely, any samples that do not meet these criteria are classified as negatives.

Contrastive learning employs a loss function that utilizes similarity as a measure to bring positive samples closer to each other while separating negative samples. Formally, the loss function for a pair of samples p_i^w, p_i^d can be formulated as follows:

$$L_i^{w,d} = -\log \frac{\exp(\text{sim}(p_i^w, p_i^d)/\tau)}{\sum_{j=1}^T \mathbb{1}_{[i \neq j]} \exp(\text{sim}(p_i^w, p_j^d)/\tau)} \quad (6)$$

where $\mathbb{1}_{[i \neq j]} \in \{0, 1\}$, it set to 1 iff $i \neq j$. i and j are the sample indices in current batch. τ is the temperature scalar. And $\text{sim}(p_i^w, p_j^d)$ represents cosine similarity computed by

$$\text{sim}(p_i^w, p_j^d) = \frac{p_i^w \cdot p_j^d}{\|p_i^w\| \|p_j^d\|} \quad (7)$$

By iterating over all sample time points within the batch, we can calculate the final contrastive loss,

$$L_{CL} = \sum_{i=1}^T L_i^{w,d} \quad (8)$$

Total loss. Due to the low signal-to-noise ratio of EEG as a physiological signal, we begin by conducting wet and dry electrode EEG emotion classification tasks. This crucial step ensures that the features extracted by the encoders from both wet and dry electrodes effectively contribute to emotion recognition. By doing so, we establish a foundation for feature alignment steps within the contrastive learning framework.

To consolidate the aforementioned information, we employ a joint training mechanism and update our final loss function as follows:

$$L = L_W + L_D + L_{CL} \quad (9)$$

where L_W and L_D are the classification losses of wet electrode and dry electrode EEG signals respectively.

V. EXPERIMENT RESULTS

In this section, we first present the baseline emotion recognition results derived from PadWEED. Following this, we list the performance evaluation results of DECAN in comparison to other established solutions. Moreover, we conduct ablation experiments to analyze the influence of the contrastive learning

framework on classification performance within a specific configuration and visualize the embeddings produced by our model. Lastly, we study the generalization performance of DECAN through alignment experiments across subjects and datasets (SEED V), where dry and wet electrodes come from different subjects and datasets respectively.

A. Datasets

PaDWEED. The PaDWEED dataset comprises two sub-datasets: (1) The dry electrode sub-dataset includes 24-channel EEG data gathered from 16 subjects. (2) The wet electrode sub-dataset encompasses 64-channel EEG data obtained from the identical group of 16 subjects exposed to the same video stimuli. This paired cross-device dataset facilitates the research of feature alignment between dry and wet electrode recordings.

SEED V. The SEED V dataset [44] comprises 62-channel wet electrode EEG recordings from 20 subjects, with each subject participating in three sessions. During each session, there are 15 trials, corresponding to 15 movie clips evenly distributed across 5 emotional states (happy, sad, disgust, fear, and neutral).

B. Implementation details

To evaluate the performance of our emotion recognition results and establish baseline classification results for the proposed database, we perform individual participant classification using well-established approaches commonly employed in prior research on affective computing. Specifically, we employ a support vector machine (SVM) classifier with a linear kernel, logistic regression (LR), and DNN. These methods have demonstrated their effectiveness in previous studies and serve as reference models in our analysis [38], [45]–[47]. Specifically, regarding the LR model, we utilize the default function provided by the scikit-learn module. As for the SVM classifier, we employ the function available in the scikit-learn module with a linear kernel. To determine the optimal misclassification cost parameter C for the linear SVM, we conduct a grid search over the range of $[2^{-10}, 2^{-9}, \dots, 2^{10}]$ and $[0.1, 20]$, using a step size of 0.5 for the large-step and small-step scenarios, respectively. In the DNN model utilized in this paper, we integrate three hidden layers with 128, 64, and 32 hidden units for the wet electrode system, whereas the dry electrode system utilized only the last two hidden layers. The output layer comprised five units, each corresponding to one of the five emotions mentioned above.

In our quest to optimize the model, we empirically adjust to specific hyperparameters based on preliminary results. To enhance the training process, we employ the RMSprop optimizer with the learning rate selected from $[1^{-4}, 9^{-4}], [1^{-3}, 9^{-3}]$ and $[1^{-2}, 9^{-2}]$ using step size of $2^{-4}, 2^{-3}, 2^{-2}$ respectively for optimization. For all the experiments conducted, our model is trained for a maximum of 15,000 epochs. In addition, the sample length of 5 seconds is used to split the entire trial into segments, which facilitates consistent evaluation and benchmarking against the baseline performance of the dataset.

To strike a balance between the effectiveness of the contrastive learning component and the overall stability of the

training process for the DECAN model, we make a strategic decision regarding the contrastive learning temperature scalar. Specifically, we selected a value of 0.5 as a compromise. In our efforts to enhance the quality of the projection of representations into the latent space, we conduct tuning of the number of hidden units within the projection head. Specifically, we explore a range of options, including 64, 128, 256, and 512 hidden units.

C. Baseline results of PaDWEED

Conditioned on intra-subject. To validate the classification performance in the condition of intra-subject, we employ a leave-one-block-out (LOBO) cross-validation technique. In each step of the cross-validation process, one block of samples is held out as the test set, while the classifier is trained on the samples from the remaining blocks. This process is repeated for all blocks' data, ensuring that each sample is used for testing exactly once. Fig.4 presents the results of three different classifiers, LR, SVM and DNN in wet and dry electrode systems. Specific quantitative results are shown in Table II where the best performance in each case is highlighted in bold.

In an effort to offer valuable insights into optimizing the recognition system for enhanced performance, a paired t-test is conducted on the recognition results of the three models, focusing on performance disparities in both the wet and dry electrode systems. The statistical test results presented in Fig 4 (a) and (b) demonstrate significant differences among the models. In the wet electrode system, the DNN model exhibits superior performance compared to both the SVM and LR models, with $p \leq 0.001$. Similarly, in the dry electrode system, the DNN model achieved the best performance among the three models. However, one notable distinction from the wet electrode system is that the DNN model exhibited a more pronounced significance difference compared to the SVM model, with $p \leq 0.001$. These findings highlight the effectiveness of the DNN model in both electrode systems and emphasize its superiority over alternative models.

Conditioned on inter-subject. We employ the Leave-One-Subject-Out (LOSO) cross-validation approach. This validation scheme involved utilizing data from 15 participants for training purposes, while the data from one participant is set aside for testing. The accuracy results for all subjects are depicted in Fig.4 (c) and (d), showcasing a comprehensive overview of the findings. A notable observation is the substantial decrease in recognition accuracy for all three models when compared to the intra-subject scenario, regardless of whether the EEG system utilizes dry or wet electrodes. Within the

wet electrode EEG system, both the DNN and SVM models achieve significantly superior results compared to the LR model. However, no significant difference is observed between the DNN and SVM models. Conversely, in the dry electrode EEG system, the DNN model consistently outperforms the SVM and LR models. This difference in performance is statistically significant, with $p \leq 0.001$ for both the SVM and LR models.

Baseline results comparison. Based on our analysis, we found a statistically significant difference in recognition accuracy between the wet electrode system and the dry electrode system in both intra-subject and inter-subject scenarios ($p \leq 0.001$). Average accuracy that is 14.53% higher in intra-subject and 9.20% higher in inter-subject compared to the dry electrode system. The superior performance of the wet electrode system can be ascribed to its ability to mitigate signal noise more effectively, thereby providing more valuable information compared to the dry electrode system. Additionally, in inter-subject experiment, the wet electrode system demonstrates a decrease in emotion recognition accuracy by 16.33%, while the dry electrode system exhibits an 24.43% decrease compared to intra-subject which may be attributed to the increased variability when combining data from all subjects and differences in distribution between source and target domains.

D. Performance Evaluation of DECAN

Main results. We consider 5 baseline methods, all originally designed for wet-electrode EEG-based emotion recognition rather than dry-electrode systems, so we adjust the settings for these methods and ensured comparability of the results. According to Table IV, DECAN consistently demonstrates superior performance across all metrics. Specifically, CDRC achieves the highest performance with an accuracy of 55.01% and an F1 score of 48.59%, showing a margin of 3.57% and 4.52% over the second-best baseline method DNN. These enhancements suggest that DECAN can effectively incorporate wet electrode information into dry electrode EEG-based emotion recognition systems and hold great promise for enhancing the accuracy and robustness of such systems.

Fig.5 compares the classification confusion matrices of the DNN (suboptimal method) with the proposed DECAN model in the intra-subject experiment. The experiment results indicate that our model achieves higher classification accuracy for almost all emotion categories, particularly excelling in the recognition of anger among negative emotions.

TABLE II

LOBO TEST RESULTS OF BASED ON THREE DIFFERENT CLASSIFIERS. THE MEAN AND STANDARD DEVIATION (%) OF ACCURACIES ARE SHOWN.

Type	LR [47]		SVM [45]		DNN [46]	
	Acc	Std	Acc	Std	Acc	Std
Wet	54.39	7.61	59.62	5.91	68.74	9.38
Dry	30.73	7.98	41.24	6.81	51.44	5.31

TABLE III

LOSO TEST RESULTS OF BASED ON THREE DIFFERENT CLASSIFIERS. THE MEAN AND STANDARD DEVIATION (%) OF ACCURACIES ARE SHOWN.

Type	LR [47]		SVM [45]		DNN [46]	
	Acc	Std	Acc	Std	Acc	Std
Wet	34.87	8.71	40.43	10.82	44.31	5.60
Dry	22.53	5.41	27.52	3.87	35.11	2.11

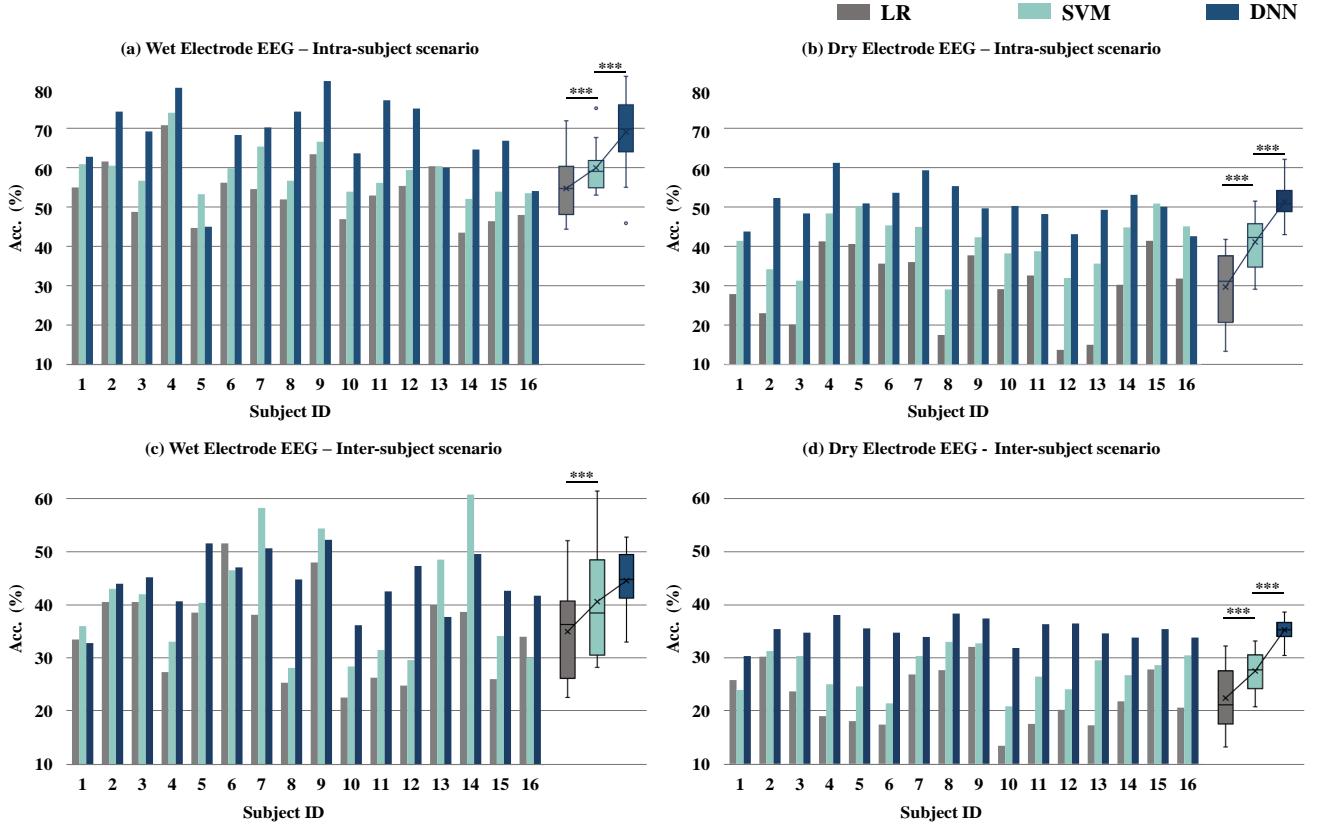


Fig. 4. The baseline results of the wet-electrode and dry-electrode based emotion classification in the condition of both intra-subject and inter-subject scenarios on our proposed dataset.

Ablation studies. We set the framework without the primary component contrastive learning as the baseline method in this experiment. The experimental results of the intra-subject scheme are depicted in Fig.6. Upon analysis, it is evident that the emotion recognition accuracy of the dry electrode EEG system exhibits improvement for 13 out of the 16 subjects, with subject 5 and 16 showing enhancements of up to 13.75% and 17.07% respectively. We have further applied the paired t-test to find whether there are significant differences between the results of the baseline approach and our model at the significant level as 0.05. In our analysis, the calculated p-value was determined to be 0.01. This result suggests that there is indeed a significant difference between the results of the baseline approach and our model, demonstrating the effectiveness of the contrastive learning module.

We further investigate the effectiveness of introducing the contrastive learning module in our proposed DECAN by visualizing the latent EEG patterns. The feature distributions of three subjects from the PaDWEED dataset are presented in a two-dimensional space using t-SNE. For each subject, two latent features with a length of 5 seconds are randomly visualized, corresponding to the two methods from a trial in the testing set. The results, as depicted in Fig.7, clearly demonstrate that our approach, which combines the base emotion model with the contrastive learning strategy, yields more effective dry electrode EEG features.

Generalization test on inter-subject intra-dataset sce-

nario. In this experiment conducted on the PaDWEED dataset, our objective is to assess the performance of DECAN in dry electrode emotion recognition when the wet electrode signals paired with the dry electrode ones originated from different subjects, introducing inter-subject variations. Specifically, we utilize "one-to-one" strategy, where the wet electrode EEG data from one subject are sequentially paired with the dry electrode EEG data from other subjects to construct the training set. Following this, the remaining dry electrode data trials from the subjects who provide dry electrode data in the training set are employed as the test set in sequence. For instance, we utilize the wet electrode EEG data from subject 4 as the training data, and the corresponding experiment results are illustrated in Fig.8 (a). It is evident that leveraging the same wet electrode EEG data has enhanced the accuracy for 13 out of 15 subjects, with the average accuracy showing a relative improvement of 5.99%, indicating DECAN can effectively mitigate the individual differences among the subjects.

Generalization test on inter-subject inter-dataset scenario. In this experiment, we additionally take into account the variations between datasets in addition to the individual variances among participants. To achieve inter-dataset pairing, the wet electrode EEG signals in the training data are sourced from the SEED V dataset (subject 1), while the dry electrode signals are obtained from the PaDWEED dataset. We specifically focused on cases with identical emotion categories (such as happy, neutral, sad, and fearful emotions) in both datasets,

TABLE IV

COMPARISON WITH REPRESENTATIVE METHODS THAT ARE WIDELY USED IN THE EEG ANALYSIS FIELD. THE MEAN AND STANDARD DEVIATION (%) OF EVALUATION METRICS ARE SHOWN.

Ref.	Model	Accuracy	Precision	Recall	F1	AUROC	AUPRC
[47]	LR*	30.73±7.98	27.17±8.89	31.03±7.62	26.24±7.10	—	—
[45]	SVM*	41.24±6.81	35.90±7.66	41.16±6.86	34.54±6.49	—	—
[48]	DGCNN*	43.37±8.49	34.44±11.69	42.45±10.01	34.16±10.45	64.73±11.55	44.43±9.04
[49]	Conformer*	39.41±4.42	39.11±6.18	39.36±4.45	35.24±5.42	64.06±6.02	39.32±5.99
[46]	DNN*	51.44±5.31	46.95±8.24	51.29±5.24	44.07±6.34	67.43±5.28	43.59±5.54
	DECAN (Ours)	55.01±5.07	51.64±8.46	54.91±5.01	48.59±6.27	69.45±5.73	45.2±6.07

* indicates the results are obtained by our own implementation. Best results are in bold, while the second-best results are underlined.

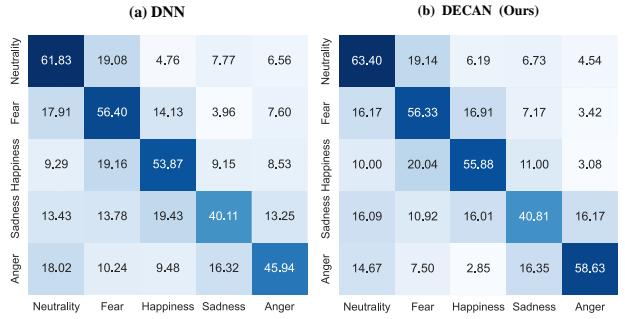


Fig. 5. Confusion matrices of DNN (the second-best method) and the DECAN model for dry electrode EEG-based emotion recognition on PaDWEED dataset.

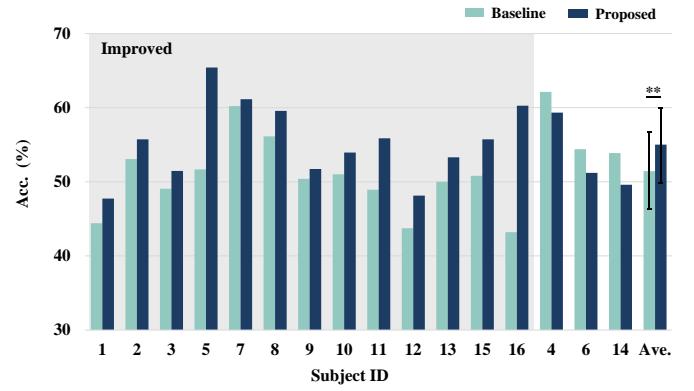


Fig. 6. Emotion recognition performance of our DECAN via ablation study on our proposed dataset in the intra-subject scheme.

utilizing preprocessed 4-second EEG signals for consistency. Corresponding results are shown in Fig. 8 (b), it demonstrates that DECAN can enhance the accuracy of dry electrode EEG emotion recognition for the majority of subjects, with an average relative improvement of 5.14%. This enhancement indicates that DECAN can address the challenges posed by cross-dataset scenarios and effectively boost the performance of dry electrode emotion recognition utilizing non-homologous wet electrode datasets.

VI. DISCUSSION

Historically, there have been two mainstreams regarding improving the emotion recognition performance of dry electrode systems, refining prototyping workflows [16], [18], [20], [21] and optimizing hardware [17]. Here, to address the limitation of challenging hardware advancements in the short term, we focused on the former one and proposed a contrastive learning-based method to enhance the dry electrode EEG data emotion recognition performance. Unlike conventional methods that primarily focused on designing encoders to extract efficient EEG features from dry electrode systems [16], [18], [20], which often heavily depended on the quality of dry electrode EEG signals, our approach takes into consideration the challenging aspect of extracting efficient features from dry electrode EEG signals with relatively low signal-to-noise ratio. To overcome this challenge, we proposed to leverage the advantages offered by wet electrode EEG systems to enhance the emotion recognition capabilities of dry electrode systems.

The intra-subject analysis on the PaDWEED dataset demonstrates that our proposed DECAN effectively improves the

emotion recognition performance of dry electrode EEG systems by leveraging knowledge obtained from wet electrode EEG systems. This finding supports the existence of shared features between dry and wet electrodes when performing the same task, which is consistent with previous research [14]–[16]. It also highlights the potential of contrastive learning methods in uncovering this specific information. Moreover, the inter-subject feature alignment analysis reveal promising results in improving the emotion recognition performance of an individual's dry electrode EEG by utilizing wet electrode

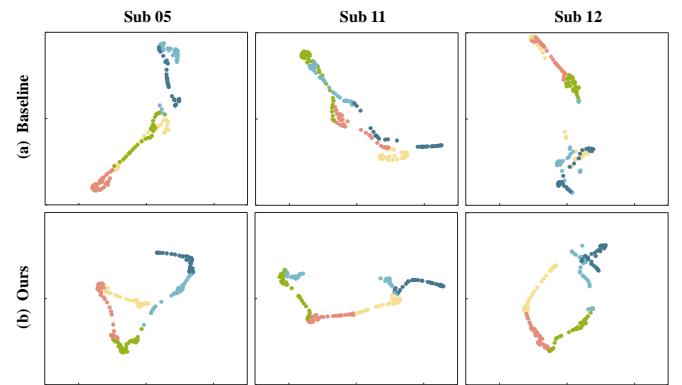


Fig. 7. Visualization of latent features using t-SNE on the PaDWEED dataset. We presented the features extracted by two models: (a) Baseline, (b) DECAN. The different colors represent different emotions.

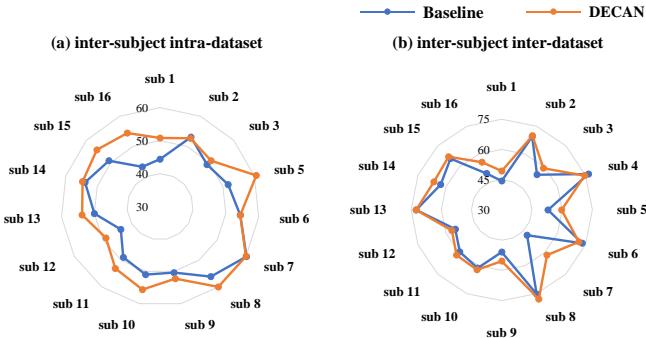


Fig. 8. Generation tests of DECAN. (a) The results of dry electrode EEG emotion recognition performance of the remaining subjects assisted by wet electrode EEG signals from subject 4. (b) The results of dry electrode EEG emotion recognition performance of the subjects in PaDWEED dataset assisted by wet electrode EEG signals from SEED V dataset (subject 1).

EEG data from other subjects and even from non-homologous dataset. This suggests that DECAN proposed in this study can effectively align EEG signals across devices and subjects. Consequently, it eliminates the need for each subject to undergo the corresponding wet electrode experiment, significantly reducing the complexity of the overall experimental setup.

The contrastive learning architecture in DECAN has been shown to be effective in improving dry electrode EEG emotion recognition performance. Here, we further investigate its impact on the accuracy obtained with different frequency sub-band signals for each of the 16 participants and visualize the average results in Fig. 9. In the baseline results, the recognition accuracy is notably higher in the high-frequency bands, especially the gamma band, which is consistency with previous studies [50]–[52]. Upon employing the DECAN model with the contrastive learning module, enhancements are observed in the recognition accuracy of the full-band signal. Moreover, there has been a significant improvement in the recognition accuracy of the delta and beta bands, while this outcome of other sub-bands has remained stable. This suggests that delta and beta bands also play a substantial role in dry electrode EEG emotion recognition, aligning with findings from prior research [53].

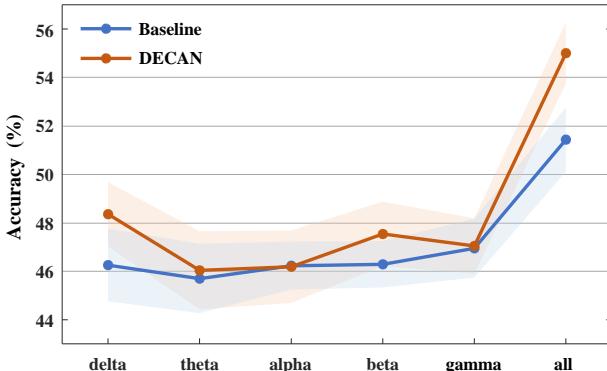


Fig. 9. Emotion recognition performance of our DECAN via ablation study on our proposed dataset in the intra-subject scheme.

Lastly, to our knowledge, PaDWEED is the first database which contains paired dry and wet electrode EEG data from collected with the same subjects using the same set of video stimuli and experimental protocol. This database can serve as a valuable research resource for scholars in the field of affective computing. It offers opportunities to investigate emotion recognition and the underlying mechanisms of both wet electrode and dry electrode systems. Additionally, it enables the exploration of the relationship between physiological signal patterns captured by these two types of electrodes. Moreover, the database goes beyond single modality research, as it includes simultaneous collection of peripheral physiological signals such as ECG, EOG, GSR, BVP, RSP and SKT signals. This comprehensive dataset can provide a foundation for conducting research on multimodal emotion recognition, whether using dry electrodes or wet electrodes. By incorporating multiple physiological signals, researchers can further enhance the accuracy and robustness of emotion recognition systems.

Indeed, it must be acknowledged that the achieved performance in dry electrode emotion recognition within this study still leaves ample room for enhancement, which may be attributed to the fact that the current encoder architectures may not be able to fully explore the entire range of EEG features, thereby constraining their capacity to extract the valuable information embedded in wet EEG data effectively. To address this issue, future investigations could concentrate on refining the architecture and configuration of encoders optimized for extracting wet electrode EEG features, which may involve exploring more advanced and complex neural network architectures.

VII. CONCLUSION

In this study, we introduce a denoising encoder via contrastive learning alignment network (DECAN) to enhance the performance of dry electrode EEG emotion recognition with the help of wet electrode EEG signals: (1) We propose the DECAN model, a model consisting of two partially shared DNN models and a feature alignment contrastive learning strategy to extract efficient dry electrode EEG emotion features. (2) We construct a new dataset named PaDWEED to better support our research, which contains paired dry and wet electrode EEG datasets from 16 subjects, along with peripheral physiological signals such as ECG, EOG, RSP, GSR, BVP, and SKT. (3) Experimental results on the PaDWEED dataset demonstrate that our proposed model achieves state-of-the-art performance on the dry electrode EEG emotion recognition task. Additionally, experiments involving feature alignment across subjects and datasets reveal that DECAN can effectively overcome more challenging scenarios.

REFERENCES

- [1] K. Ezzamel and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Inf. Fusion*, vol. 99, p. 101847, 2023.
- [2] Y. Peng, H. Liu, W. Kong, F. Nie, B. Lu, and A. Cichocki, "Joint EEG feature transfer and semisupervised cross-subject emotion recognition," *IEEE Trans. Ind. Informatics*, vol. 19, no. 7, pp. 8104–8115, 2023.

- [3] S. Xiao, X. Qiu, C. Tang, and Z. Huang, "A spatial-temporal ECG emotion recognition model based on dynamic feature fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5.
- [4] K. Kipli, A. A. A. Latip, K. Lias, N. Bateni, S. M. Yusoff, N. M. A. Tajudin, M. Jalil, K. Ray, M. Shamim Kaiser, and M. Mahmud, "Gsr signals features extraction for emotion recognition," in *Proceedings of Trends in Electronics and Health Informatics: TEHI 2021*. Springer, 2022, pp. 329–338.
- [5] L. Mirmohamadsadeghi, A. Yazdani, and J. Vesin, "Using cardio-respiratory signals to recognize emotions elicited by watching music video clips," in *18th IEEE International Workshop on Multimedia Signal Processing, MMSP 2016, Montreal, QC, Canada, September 21-23, 2016*. IEEE, 2016, pp. 1–5.
- [6] Y. Ding, X. Hu, Z. Xia, Y. Liu, and D. Zhang, "Inter-brain EEG feature extraction and analysis for continuous implicit emotion tagging during video watching," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 92–102, 2021.
- [7] B. García-Martínez, A. Martínez-Rodrigo, R. Alcaraz, and A. Fernández-Caballero, "A review on nonlinear methods using electroencephalographic recordings for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 801–820, 2021.
- [8] J. Yin, S. Wang, T. Tat, and J. Chen, "Motion artefact management for soft bioelectronics," *Nature Reviews Bioengineering*, pp. 1–18, 2024.
- [9] C. He, Y. Chen, C. Phang, C. Stevenson, I. Chen, T. Jung, and L. Ko, "Diversity and suitability of the state-of-the-art wearable and wireless EEG systems review," *IEEE J. Biomed. Health Informatics*, vol. 27, no. 8, pp. 3830–3843, 2023.
- [10] F. Freire, M. Becchi, S. Ponti, E. Miraldi, and A. Strigazzi, "Impedance spectroscopy of conductive commercial hydrogels for electromyography and electroencephalography," *Physiological Measurement*, vol. 31, no. 10, p. S157, 2010.
- [11] Y. Peng, W. Wang, W. Kong, F. Nie, B. Lu, and A. Cichocki, "Joint feature adaptation and graph adaptive label propagation for cross-subject emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1941–1958, 2022.
- [12] 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. Affect. Comput.*, vol. 14, no. 3, pp. 2496–2511, 2023.
- [13] W. Zhang, F. Wang, Y. Jiang, Z. Xu, S. Wu, and Y. Zhang, "Cross-subject eeg-based emotion recognition with deep domain confusion," in *Intelligent Robotics and Applications - 12th International Conference, ICIRA 2019, Shenyang, China, August 8-11, 2019, Proceedings, Part I*, ser. Lecture Notes in Computer Science, H. Yu, J. Liu, L. Liu, Z. Ju, Y. Liu, and D. Zhou, Eds., vol. 11740. Springer, 2019, pp. 558–570.
- [14] J. W. Y. Kam, S. Griffin, A. Shen, S. Patel, H. Hinrichs, H. Heinze, L. Y. Deouell, and R. T. Knight, "Systematic comparison between a wireless EEG system with dry electrodes and a wired EEG system with wet electrodes," *NeuroImage*, vol. 184, pp. 119–129, 2019.
- [15] H. Hinrichs, M. Scholz, A. K. Baum, J. W. Kam, R. T. Knight, and H.-J. Heinze, "Comparison between a wireless dry electrode eeg system with a conventional wired wet electrode eeg system for clinical applications," *Scientific reports*, vol. 10, no. 1, p. 5218, 2020.
- [16] Y.-T. Lan, D. Peng, W. Liu, Y. Luo, Z. Mao, W.-L. Zheng, and B.-L. Lu, "Investigating emotion eeg patterns for depression detection with attentive simple graph convolutional network," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023, pp. 1–4.
- [17] Z. Fangmeng, P. Siriaraya, D. Choi, and N. Kuwahara, "Textile eeg cap using dry-comb electrodes for emotion detection of elderly people," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020.
- [18] P. Lakhani, N. Banluesombatkul, V. Changniam, R. Dhithijaiyaratn, P. Leelaarporn, E. Boonchieng, S. Hompoonsup, and T. Wilaprasitporn, "Consumer grade brain sensing for emotion recognition," *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9896–9907, 2019.
- [19] 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 Informatics*, vol. 22, no. 1, pp. 98–107, 2018.
- [20] M. M. Javaid, M. A. Yousaf, Q. Z. Sheikh, M. M. Awais, S. Saleem, and M. Khalid, "Real-time eeg-based human emotion recognition," in *Neural Information Processing - 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, S. Arik, T. Huang, W. K. Lai, and Q. Liu, Eds., vol. 9492. Springer, 2015, pp. 182–190.
- [21] T. Xu, R. Yin, L. Shu, and X. Xu, "Emotion recognition using frontal eeg in vr affective scenes," in *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, vol. 1. IEEE, 2019, pp. 1–4.
- [22] M. M. Rahman, A. K. Sarkar, M. A. Hossain, M. S. Hossain, M. R. Islam, M. B. Hossain, J. M. W. Quinn, and M. A. Moni, "Recognition of human emotions using EEG signals: A review," *Comput. Biol. Medicine*, vol. 136, p. 104696, 2021.
- [23] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.
- [24] S. Rayatdoost, D. Rudrauf, and M. Soleymani, "Expression-guided EEG representation learning for emotion recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 3222–3226.
- [25] S. Bagherzadeh, K. Maghooli, A. Shalbaf, and A. Maghsoudi, "Emotion recognition using effective connectivity and pre-trained convolutional neural networks in eeg signals," *Cognitive Neurodynamics*, vol. 16, no. 5, pp. 1087–1106, 2022.
- [26] Y. Zhang, C. Cheng, S. Wang, and T. Xia, "Emotion recognition using heterogeneous convolutional neural networks combined with multimodal factorized bilinear pooling," *Biomed. Signal Process. Control.*, vol. 77, p. 103877, 2022.
- [27] S. Koelstra, C. Mühl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.
- [28] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12 177–12 191, 2019.
- [29] W. Zheng and B. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 3, pp. 162–175, 2015.
- [30] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schäffer, and F. Schwender, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *CoRR*, vol. abs/1812.02782, 2018.
- [31] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, 2016.
- [32] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, 2018.
- [33] X. Wang, S. Zhao, Y. Pei, Z. Luo, L. Xie, Y. Yan, and E. Yin, "The increasing instance of negative emotion reduce the performance of emotion recognition," *Frontiers in Human Neuroscience*, vol. 17, 2023.
- [34] W. Zheng, B. Dong, and B. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014, Chicago, IL, USA, August 26-30, 2014*. IEEE, 2014, pp. 5040–5043.
- [35] S. Wu, M. Schaefer, W. Zheng, B. Lu, and H. Yokoi, "Neural patterns between chinese and german for eeg-based emotion recognition," in *8th International IEEE/EMBS Conference on Neural Engineering, NER 2017, Shanghai, China, May 25-28, 2017*. IEEE, 2017, pp. 94–97.
- [36] L. A. Moctezuma, T. Abe, and M. Molinas, "Two-dimensional cnn-based distinction of human emotions from eeg channels selected by multi-objective evolutionary algorithm," *Scientific Reports*, vol. 12, no. 1, p. 3523, 2022.
- [37] L. Shi, Y. Jiao, and B. Lu, "Differential entropy feature for eeg-based vigilance estimation," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2013, Osaka, Japan, July 3-7, 2013*. IEEE, 2013, pp. 6627–6630.
- [38] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, 2019.
- [39] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and eeg to enhance emotion recognition," in *IJCAI*, vol. 15. Buenos Aires, 2015, pp. 1170–1176.
- [40] J.-Y. Zhu, W.-L. Zheng, Y. Peng, R.-N. Duan, and B.-L. Lu, "Eeg-based emotion recognition using discriminative graph regularized extreme learning machine," in *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 525–532.
- [41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever,

- “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763.
- [42] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J. King, “Decoding speech perception from non-invasive brain recordings,” *Nat. Mac. Intell.*, vol. 5, no. 10, pp. 1097–1107, 2023.
- [43] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607.
- [44] 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 Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715–729, 2021.
- [45] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, “AMIGOS: A dataset for affect, personality and mood research on individuals and groups,” *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, 2021.
- [46] W. Liu, W.-L. Zheng, Z. Li, S.-Y. Wu, L. Gan, and B.-L. Lu, “Identifying similarities and differences in emotion recognition with eeg and eye movements among chinese, german, and french people,” *Journal of Neural Engineering*, vol. 19, no. 2, p. 026012, 2022.
- [47] M. Wu, W. Teng, C. Fan, S. Pei, P. Li, and Z. Lv, “An investigation of olfactory-enhanced video on eeg-based emotion recognition,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1602–1613, 2023.
- [48] 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.
- [49] Y. Song, Q. Zheng, B. Liu, and X. Gao, “Eeg conformer: Convolutional transformer for eeg decoding and visualization,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.
- [50] C. Chen, Z. Li, F. Wan, L. Xu, A. Bezerianos, and H. Wang, “Fusing frequency-domain features and brain connectivity features for cross-subject emotion recognition,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.
- [51] Y. Peng, W. Kong, F. Qin, F. Nie, J. Fang, B.-L. Lu, and A. Cicchetti, “Self-weighted semi-supervised classification for joint eeg-based emotion recognition and affective activation patterns mining,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.
- [52] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, “Sparsedgcn: Recognizing emotion from multichannel eeg signals,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 537–548, 2023.
- [53] Y. Zhang, G. Yan, W. Chang, W. Huang, and Y. Yuan, “Eeg-based multi-frequency band functional connectivity analysis and the application of spatio-temporal features in emotion recognition,” *Biomedical Signal Processing and Control*, vol. 79, p. 104157, 2023.

Complex Emotion Recognition System using basic emotions via Facial Expression, EEG, and ECG Signals: a review

Javad Hassannataj Joloudari ^{1,2,3}, Mohammad Maftoun ⁴, Bahareh Nakisa ⁵, Roohallah Alizadehsani ⁶, Meisam Yadollahzadeh-Tabari ²

¹ Department of Computer Engineering, Faculty of Engineering, University of Birjand, Iran

² Department of Computer Engineering, Babol Branch, Islamic Azad University, Babol, Iran

³ Department of Computer Engineering, Technical and Vocational University (TVU), Tehran 4631964198, Iran

⁴ Department of Artificial Intelligence, Technical and Engineering Faculty, South Tehran Branch, Islamic Azad University, Tehran, Iran

⁵ School of Information Technology, Faculty of Science Engineering and Built Environment, Deakin University, Geelong, Vic, Australia

⁶ Institute for Intelligent Systems Research and Innovation (IISRI) Deakin

University, Waurn Ponds, Australia

Abstract: The Complex Emotion Recognition System (CERS) deciphers complex emotional states by examining combinations of basic emotions expressed, their interconnections, and the dynamic variations. Through the utilization of advanced algorithms, CERS provides profound insights into emotional dynamics, facilitating a nuanced understanding and customized responses. The attainment of such a level of emotional recognition in machines necessitates the knowledge distillation and the comprehension of novel concepts akin to human cognition. The development of AI systems for discerning complex emotions poses a substantial challenge with significant implications for affective computing. Furthermore, obtaining a sizable dataset for CERS proves to be a daunting task due to the intricacies involved in capturing subtle emotions, necessitating specialized methods for data collection and processing. Incorporating physiological signals such as Electrocardiogram (ECG) and Electroencephalogram (EEG) can notably enhance CERS by furnishing valuable insights into the user's emotional state, enhancing the quality of datasets, and fortifying system dependability. A comprehensive literature review was conducted in this study to assess the efficacy of machine learning, deep learning, and meta-learning approaches in both basic and complex emotion recognition utilizing EEG, ECG signals, and facial expression datasets. The chosen research papers offer perspectives on potential applications, clinical implications, and results of CERSs, with the objective of promoting their acceptance and integration into clinical decision-making processes. This study highlights research gaps and challenges in understanding CERSs, encouraging further investigation by relevant studies and organizations. Lastly, the significance of meta-learning approaches in improving CERS performance and guiding future research endeavors is underscored.

Keywords: Complex emotion, Basic emotion, Physiological signals , Facial, Emotion Recognition, Meta learning

1. Introduction

Affective computing is an interdisciplinary research domain that amalgamates the fundamental tenets of psychology, computer science, and cognitive science.

The automatic recognition of emotions is a crucial component of affective computing, whereby the recognition of emotional states serves as the foundation for the computer's comprehension of emotions and subsequent reactions [1]. The field of emotion recognition, which is highly appealing, has garnered significant attention from both industry and academia. Emotion recognition has numerous potential applications, including but not limited to human-computer interaction, video gaming, and continuous monitoring of infants and patients with medical conditions such as Parkinson's disease, Alzheimer's disease, depression, falls, and so on [2].

Basic emotions are primary, universal emotional states that are widely acknowledged in various societies and typically consist of happiness, sadness, fear, anger, surprise, and disgust [3]. These emotions are believed to be innate and have distinct facial expressions and physiological reactions [4]. In contrast, complex emotions are more intricate and frequently entail a blend of basic emotions alongside higher cognitive operations, such as jealousy,

pride, and guilt [5]. Differing from basic emotions, complex emotions are shaped by personal encounters, social engagements, and societal standards [6].

In fact, emotion is a complex psychological phenomenon, characterized by distinct emotional states that manifest through various physical and physiological cues. These cues can be broadly categorized into physical expressions and physiological signals. Physical expressions include facial expressions, voice intonation, gait, and body posture. Physiological signals encompass indicators such as Electroencephalogram (EEG) and Electrocardiogram (ECG) readings, Galvanic Skin Response (GSR), and other biofeedback mechanisms. Additionally, emotions can be conveyed through written text and other communication methods, acting as mediators of our internal emotional states [1,7,8]. The recognition of emotions through physiological signals is an essential aspect of scrutinizing psychological states and advancing biofeedback-based applications. The advent of the metaverse concept and the integration of physiological signal trackers into smart devices have significantly advanced this area of investigation, making it both essential and intriguing for researchers.

Studies have shown that changes in emotional states can directly influence physiological signals, providing a valuable means of emotion detection.

The Autonomic Nervous System (ANS) and Central Nervous System (CNS) are essential in overseeing various physiological responses to emotional stimuli, like heart rate, skin conductance, and brain activity. Monitoring this regulation effectively can be done using techniques such as EEG and ECG. EEG allows for precise capture of dynamic changes in brain activity in response to emotional stimuli due to its high temporal resolution [9, 10]. In contrast, ECG provides a reliable measurement of heart rate variability, which is closely associated with emotional arousal and autonomic regulation [11, 12]. These methods together offer a comprehensive understanding of the physiological basis of emotional states, enhancing the analysis of facial expressions .

This understanding is vital for developing accurate and reliable methods for detecting and interpreting emotions through physiological signals [13, 14]. The objective of the researchers is to develop emotion recognition systems that exhibit equitable levels of accuracy and responsiveness.

The present difficulty in recognizing complex emotions stems from the limited array of physiological signals considered, which could obstruct a comprehensive understanding of the subtle emotional states that individuals undergo. Although physiological signals are crucial for emotion detection, the emphasis has frequently been confined to a restricted range of measurements that may fail to encapsulate the complexities of intricate emotions. Broadening the focus to include signals such as EEG, indicative of central nervous system activity, and ECG, which tracks heart activity influenced by the autonomic nervous system and sensitive to emotional intensity and stress, can offer a more comprehensive viewpoint. By merging these varied signals, it becomes feasible to reveal the detailed interactions between different physiological systems and complex emotional reactions, thereby improving both the precision and richness of emotion recognition while overcoming the constraints imposed by a limited signal emphasis [15].

However, given that these signals are non-linear and non-stationary, it is crucial to meticulously select appropriate features to significantly enhance the accuracy of the system [16, 17]. Another type of emotion recognition as the most effective, natural, and universal signal used by human beings to communicate their emotional states and objectives is facial expression [18]. From 1974, 55% of messages relating to emotions and beliefs are conveyed by facial expression, 7% through spoken words, and the remaining 40% are paralinguistic [19]. Numerous facial expression recognition (FER) approaches have been investigated in the fields of computer vision and machine learning to encode expression information from face representations.

Machine learning methodologies, such as Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), and K-Nearest Neighbor (KNN), are frequently utilized for the purpose of determining the presence of emotional content in facial expressions [20]. Nevertheless, the application of these methodologies extends beyond facial expressions alone, encompassing the analysis of physiological signals as well. This broader scope allows for a more comprehensive approach to the recognition of emotions. Through the integration of machine learning techniques with physiological signals (e.g., heart rate, skin conductance, brain activity) and physical cues (e.g., facial expressions), it becomes feasible to attain heightened precision and resilience in detecting emotions. This all-encompassing strategy capitalizes on the advantages offered by diverse data sources, ultimately facilitating a more profound comprehension of emotional states [21, 22 , 23].

Another variety of techniques that utilize deep architectures are known as deep learning methods, and they are outperforming traditional machine learning techniques in terms of accuracy and productivity. Deep Learning approaches, especially convolutional neural networks (CNN), which primarily depend on supervised learning using manually labeled data, have been inspected for facial emotion recognition [24]. A fixed learning algorithm that is manually created is usually used to train contemporary machine learning models from scratch for a specific task. Additionally, approaches based on deep learning have demonstrated remarkable success in numerous domains. There are, however, evident disadvantages. For instance, areas possessing massive computational power and the ability to accumulate and simulate massive quantities of data have seen the most success. Many applications where data is inherently expensive or rare or where computational resources are not available are disqualified by this [25 , 26].

In summary, the examination of basic emotions frequently involves the analysis of data gathered under controlled conditions, posing a notable obstacle. While such controlled settings enable precise evaluation of emotional reactions, they fall short in capturing the complex and varied nature of emotions encountered in real-life situations [27]. Real-world emotional responses are typically more complex and influenced by a myriad of factors, resulting in a fusion of basic and complex emotional states that defy easy classification. This disparity underscores a significant limitation: the datasets utilized for training deep learning models do not fully represent the emotional diversity encountered in daily life [28]. Moreover, deep learning models, despite their effectiveness, exhibit inherent constraints within this context. They often necessitate extensive sets of labeled data, which prove challenging to acquire for complex emotions given their subjective essence and the complexities involved in accurately labeling them [29]. This limitation hampers the models' capacity to generalize and excel in practical applications, where emotions are less predictable and more reliant on context. To tackle these hurdles, advanced deep learning methods and meta-learning strategies are under exploration. These approaches strive to enhance the flexibility and resilience of emotion recognition systems by utilizing smaller, more varied datasets and integrating contextually-aware learning mechanisms [30].

On the other hand, most famous datasets such as MultiPie, SFEW, RaFD, JAFFE, FER2013, etc are related to basic emotions while emotions are frequently more complicated than what basic emotions can communicate [31]. Thus complex emotions make keeping up with current emotion recognition systems difficult [32]. For complex emotion recognition systems due to the complexity and smallness of the data set, meta-learning approaches can play a vital role. The process of distilling the experience of multiple learning episodes often covering a distribution of related tasks and using this experience to improve future learning performance is called meta-learning [33, 34]. In recent years, few-shot learning [35], continual learning [36], label noises [37], and reinforcement learning [38] as the applications of meta-learning are commonly used in complex emotion recognition. The main goals of this paper are as follows:

This research delves into the enhancement of emotion detection through the amalgamation of facial expressions, EEG, and ECG signals using meta-learning techniques: The goal is to outperform conventional machine and deep learning approaches, thereby improving the adaptability and precision of emotion recognition systems for contextually aware real-world applications.

Pioneering the Study of Complex Emotion Recognition via AI: This paper marks the first in-depth investigation dedicated to understanding and developing AI systems specifically for complex emotion recognition, setting it apart from prior research that predominantly focused on basic emotions.

Examining the Contrasts Between Basic and Complex Emotions: We are delving into the fundamental discrepancies between basic emotions, which are universally acknowledged and simple, and complex emotions, which are intricate and influenced by the situation.

Providing Details About Complex Datasets: We offer a comprehensive overview of the datasets available for studying complex emotions, highlighting the challenges and intricacies involved in their use.

Assessing Existing Research on Basic Emotion Detection: By evaluating the body of research on machine learning and deep learning techniques used for basic emotion detection, we lay the groundwork for understanding the advancements and gaps that exist.

Categorizing Concepts for Complex Emotion Recognition through Meta-Learning: In this study, we categorize various approaches to complex emotion recognition based on meta-learning techniques, highlighting key methodologies that enhance the detection and understanding of emotions while expanding the capabilities of current AI technologies.

1.1 Publication analysis and search results

In this review paper, we examined 891 studies from various databases including IEEE, Science Direct, Springer, Wiley, and Google Scholar. This collection comprised 82 papers from IEEE, 60 from Springer, 158 from Science Direct, 139 from Wiley, and 450 from Google Scholar. In the initial phase, 801 records were eliminated due to duplication and irrelevance. Consequently, 90 records proceeded to the screening stage. During screening, 29 papers were excluded for being off-topic, leaving 61 publications for further review. At the eligibility stage, 23 more papers were removed. Ultimately, 38 research papers were selected as the final set of featured studies. The process of obtaining them based on PRISMA guidelines is shown in Fig. 1.

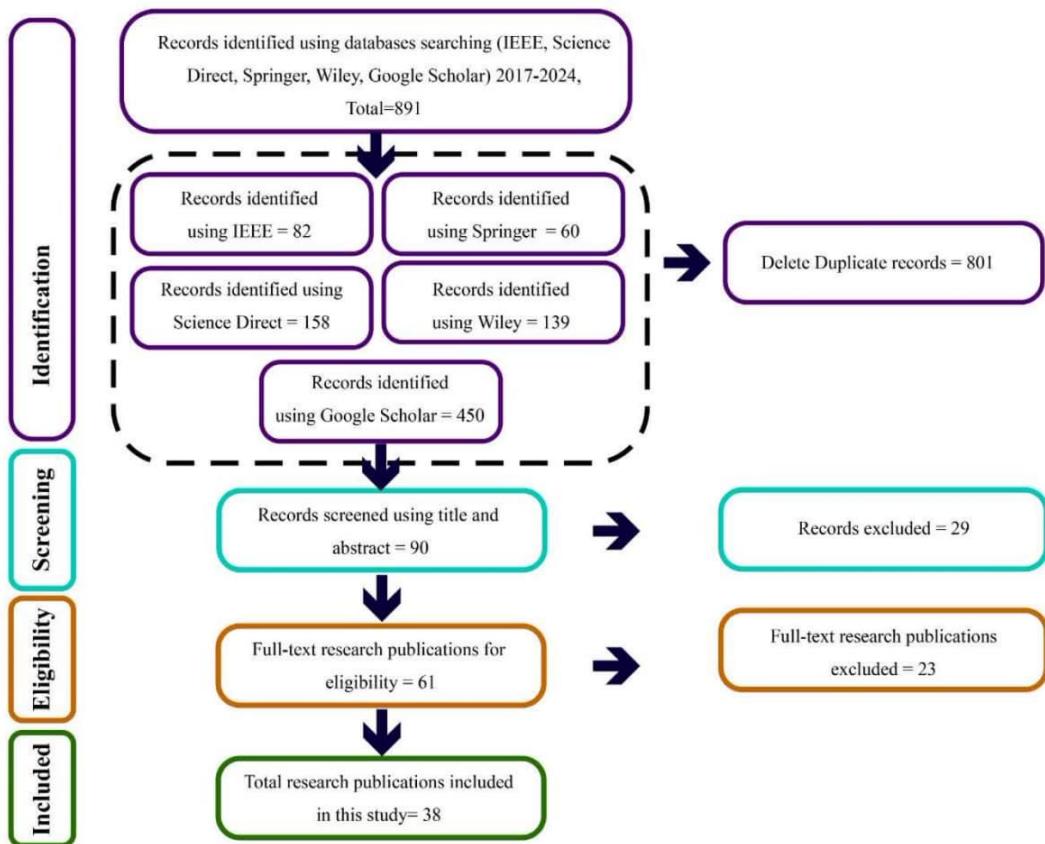


Figure 1. PRISMA flow diagram illustrating the selection process for relevant studies.

1.2 Data mapping of included studies

In this section, we provide the research papers selected from various databases. The trend chart for the number of publications on complex emotion recognition from 2017 to 2024 reveals a dynamic pattern as shown in Fig 2.

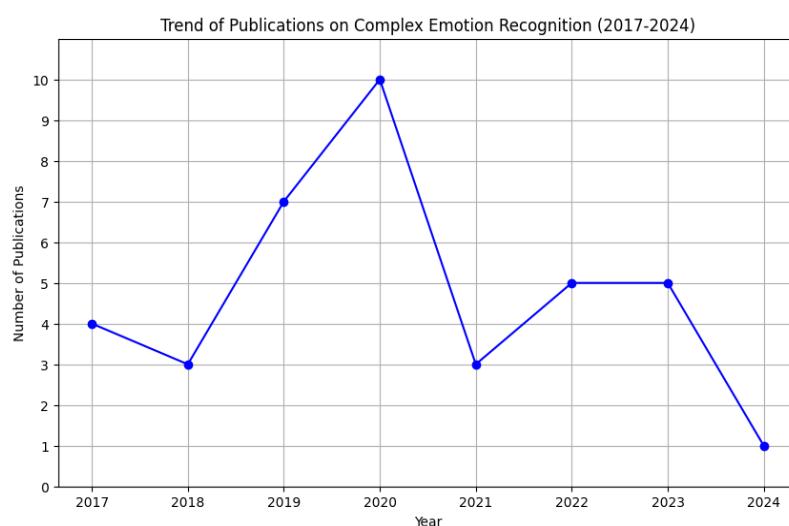


Figure 2. The trend of complex emotion recognition between 2017 and 2024.

From 2017 to 2020, there was a clear upward trend, with publications increasing from 4 to a peak of 10, indicating rising interest in the field. However, 2021 saw a sharp decline to 3 publications, possibly due to a temporary shift in research focus or resources. The numbers rebounded slightly in 2022 and 2023 with 5 publications each year, reflecting renewed efforts and interest. By 2024, the number dropped to its lowest at 1 publication, suggesting a potential winding down of major projects or shifting research priorities. Overall, the data highlights fluctuating engagement in complex emotion recognition research over the years.

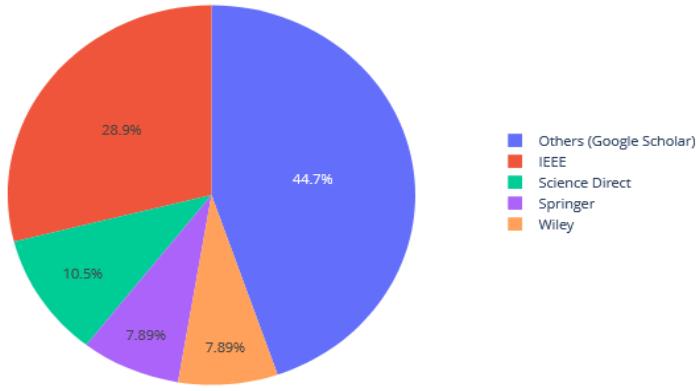


Figure 3. The allocated portion of each database in the complex emotion recognition topic.

According to Fig 3, the Google Scholar (others) database houses the largest proportion of research papers in this field, accounting for nearly half (44.7%) of the total publications. The IEEE database follows, with 30.9% of the publications. Additionally, 10.5% of relevant studies were found in Science Direct. Lastly, while the Wiley and Springer databases each contributed only 7.89% of the research papers, they have nonetheless published valuable studies on complex emotion recognition systems.

The next step of this study is structured as follows: Section 2 describes the type of emotions, the differences between them and emotion models In section 3, the complex emotion databases are reviewed and analyzed. Section 4 the preprocessing steps of developing a complex emotion recognition system are presented. Section 5 focuses on the concepts and technologies of emotion recognition systems and also limitations of traditional methods. In section 6, a review of related works are conducted. Section 7, the discussion and evaluation metrics are delivered. In Section 8, and last section are addressed into Open Research Challenges , Conclusions and future research directions.

2. Type of emotions, differences and emotion models

2.1 Basic Emotions

Emotions that are basic are brief emotional reactions that often occur automatically and have counterparts in other vertebrates. They are generally recognized worldwide and are commonly linked to specific physical reactions and facial expressions. Fear, anger, disgust, sadness, joy, and surprise are among the most frequently acknowledged basic emotions [39]. Key features of basic emotions comprise:

1. **Automaticity:** Basic emotions are activated automatically in reaction to particular stimuli without conscious deliberation. For instance, the sight of a predator may trigger fear, while the presence of a loved one may trigger joy.
2. **Distinctive Facial Expressions:** Each basic emotion is tied to a distinct facial expression that is easily recognized across diverse cultures. For instance, a frown and squinting eyes are typically associated with anger, while a grin signifies happiness. Alongside facial signals, physiological metrics like EEG and ECG offer a more profound insight into emotional experiences. The EEG method captures the brain's activities tied to various emotional states, while the ECG technique evaluates heart rate variability and autonomic nervous system responses, both being notably affected by emotional excitement and stress. Collectively, these indicators present a more comprehensive view of emotion, merging external signals with internal physiological reactions.

3. **Physiological Responses:** Basic emotions come with specific physiological alterations in the body, such as variations in heart rate, breathing, and hormone levels. Fear, for example, initiates the "fight or flight" response, resulting in heightened heart rate and adrenaline release.
4. **Evolutionary Roots:** Basic emotions are thought to have distinct evolutionary functions tied to survival and procreation. Fear, for instance, aids individuals in reacting to dangers, while joy strengthens social connections and motivates actions that enhance overall well-being.

2.2 Complex Emotions

Complex emotions are intricate and multifaceted emotional experiences that encompass a blend of basic emotions, cognitive processes, and social influences. Unlike basic emotions, which are generally widespread, complex emotions can vary significantly among individuals and societies [40]. Here are the primary characteristics:

1. **Cognitive Components:** Complex emotions entail substantial cognitive assessment and understanding of situations. Individuals may have to evaluate the context, their personal beliefs and values, and the perspectives of others to fully understand and experience intricate feelings. For instance, feelings of jealousy may emerge from interpreting a situation as a threat to a valued relationship.
2. **Cultural and Individual Variation:** In contrast to basic emotions, which are predominantly uniform, complex emotions can differ greatly among individuals and cultures. Cultural norms, values, and personal experiences have a substantial impact on shaping the experience and expression of complex emotions. For example, the way pride is felt and exhibited may vary across cultures.
3. **Extended Duration:** Complex emotions typically endure longer than basic emotions and may transform over time as new information and experiences are assimilated. For example, feelings of guilt may persist as individuals contemplate their actions and their repercussions.
4. **Interpersonal Functions:** Complex emotions often serve a vital role in managing social connections and navigating social standards. They assist individuals in comprehending and responding to others' emotions, negotiating social hierarchies, and upholding social ties. For instance, expressions of romantic love may entail a complex interplay of emotions like affection, desire, and commitment.

Basic emotions and complex emotions differ in several key aspects. Firstly, regarding duration, basic emotions are typically short-lived and immediate, whereas complex emotions tend to persist for extended periods. Secondly, in terms of cognitive involvement, basic emotions often arise automatically with minimal cognitive processing, while complex emotions require significant cognitive appraisal and interpretation of situations. Thirdly, basic emotions have clear evolutionary functions primarily related to survival, whereas complex emotions have evolved to manage more sophisticated social interactions and relationships. Finally, basic emotions are universally recognized and expressed similarly across cultures, whereas complex emotions can vary significantly based on cultural and individual differences in expression and recognition [41,42,43].

2.3 Differences Between Basic and Complex Emotions

To further clarify the differences between basic and complex emotions, it is crucial to explore their cognitive and social implications more deeply. Basic emotions like joy, fear, anger, sadness, disgust, and surprise are often seen as universal and innate, arising automatically in response to stimuli with little conscious thought. These emotions are believed to have developed to fulfill specific survival purposes, such as fear activating fight-or-flight reactions to possible dangers, and disgust aiding individuals in avoiding harmful substances. Studies by [3] and others have shown that basic emotions are expressed and understood similarly across various cultures, hinting at a common evolutionary origin.

In contrast, complex emotions such as guilt, shame, envy, pride, and love involve more intricate cognitive processes, often requiring self-reflection, comparison with others, and an understanding of societal norms and expectations. These emotions are not only enduring but also dependent on context, influenced by personal encounters, cultural upbringing, and the particular social setting. For example, feeling shame involves evaluating one's actions in relation to societal standards and anticipating others' judgments, which can differ significantly among cultures [44, 45]. Research indicates that while the fundamental elements of these emotions might be universal, their expression, interpretation, and importance are molded by cultural context and individual distinctions [46].

Furthermore, complex emotions frequently result from combinations or interactions of basic emotions [47] and are more likely to involve mixed sentiments or uncertainty. For instance, jealousy could blend fear (of losing a

relationship) with anger (towards a perceived rival) and sadness (due to feeling unappreciated). This complex interplay of emotions showcases the sophisticated cognitive processes that underlie complex emotions, making them more challenging to investigate and comprehend compared to basic emotions [48]. Additionally, the influence of language in expressing and shaping complex emotions should not be underestimated; different societies possess specific terms and ideas that capture nuanced emotional states, emphasizing further the diversity in how these emotions are felt and conveyed [49].

Basic emotions and complex emotions differ in several key aspects. Firstly, regarding duration, basic emotions are typically short-lived and immediate, whereas complex emotions tend to persist for extended periods. Secondly, in terms of cognitive involvement, basic emotions often arise automatically with minimal cognitive processing, while complex emotions require significant cognitive appraisal and interpretation of situations. Thirdly, basic emotions have clear evolutionary functions primarily related to survival, whereas complex emotions have evolved to manage more sophisticated social interactions and relationships. Finally, basic emotions are universally recognized and expressed similarly across cultures, whereas complex emotions can vary significantly based on cultural and individual differences in expression and recognition [50,51,52,53,54].

2.4 Emotion models

Defining emotion, or affect, is essential in setting criteria in affective computing. The understanding of emotions was initially presented by Ekman in the 1970s. Despite psychologists' various attempts to classify emotions in fields like neuroscience, philosophy, and computer science, there is no universally agreed-upon model of emotions. However, the two main types of emotion models commonly utilized are the discrete (or categorical) emotion model and the dimensional (or continuous) emotion model [55].

In the discrete emotion modelas which illustrated by the wheel of emotions depicted in Fig 4, individuals typically choose one emotion from a set of predefined emotions that best represents their feelings. This model categorizes emotions clearly and is based on Ekman's early research.

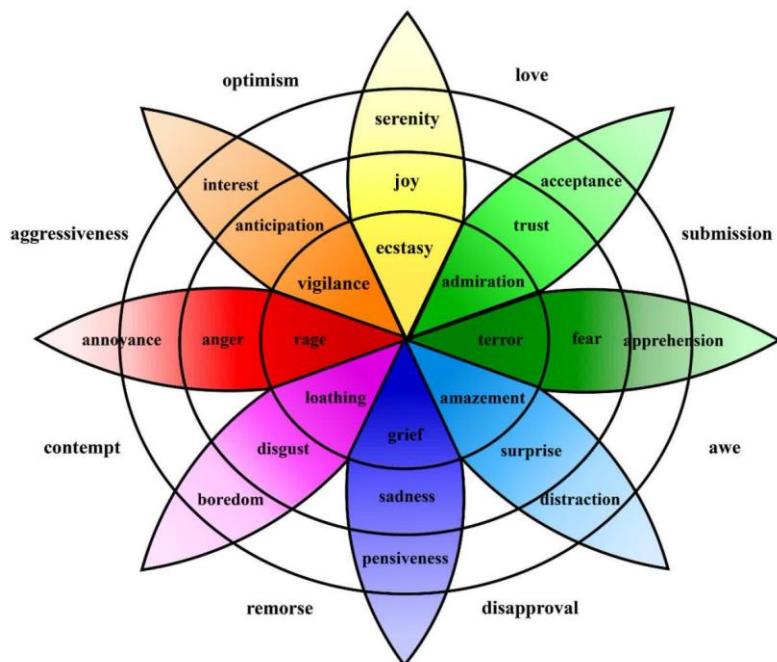


Figure 4. Wheel of emotions based on discrete models.

Conversely as shown in Fig 5, the dimensional model portrays emotions using quantitative measures across various dimensions. This method often utilizes tools such as the Self-Assessment Manikin (SAM) or Feeltrace. SAM uses images of SAMmanikins to evaluate the static level of a dimension at a specific moment, while Feeltrace monitors emotional data continuously over time [56].

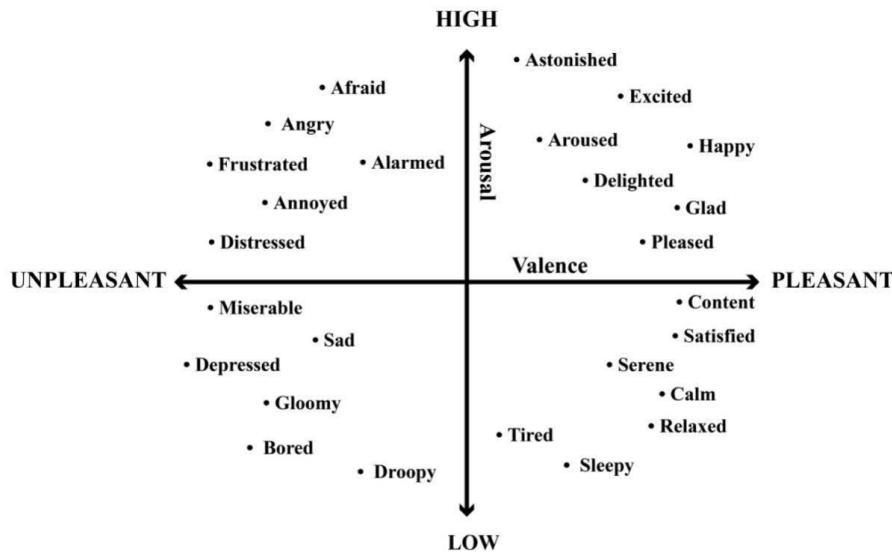


Figure 5. Emotions based on dimensional models.

Both models emphasize different facets of human emotion and provide insights into how emotions are perceived and understood by the human mind. The discrete model sorts emotions into categories with distinct labels, while the dimensional model captures the intricacy and variability of emotional states through continuous assessment. Together, these models contribute to a thorough comprehension of genuine emotional states and are crucial in the realm of affective computing.

3. Datasets used in scientific studies

Datasets are fundamental for developing every AI-based system, encompassing the essential attributes necessary for facial emotion recognition, as well as for the analysis of EEG and ECG signals. These datasets facilitate the extraction and interpretation of results, and they are combined in advance to support the development of algorithms for recognizing facial emotions and analyzing physiological signals. They include a variety of data types such as static images, videos, or a combination of both [57, 58]. Additionally, some datasets include pictures taken in staged settings using laboratory-controlled environments or are provided by psychology centers. Moreover, datasets derived organically from real-world environments are also available.

3.1 Basic and complex Emotion Datasets

CMED

Based on existing spontaneous micro-expression datasets such as CASME I, CASME II, CAS(ME), and SAMM, researchers created the Compound Micro-expression Database (CMED). These datasets mainly gather facial expression data rather than physiological signals. Due to their subtle features, it is challenging to analyze the motion track and characteristics of micro-expressions. Consequently, generating compound micro-expression images presents numerous obstacles and limitations. The data in these datasets is captured using high-speed cameras that record slight facial movements at a high frame rate, typically ranging from 100 to 200 frames per second, enabling the detection of micro-expressions lasting between 1/25th and 1/5th of a second. Participants in these datasets engage in various tasks or are exposed to specific stimuli to evoke spontaneous emotional reactions, which are then documented [59, 60, 61, 62, 63]. The number of participants varies among the datasets. For example, the CASME II dataset contains recordings from 26 individuals, while the SAMM dataset includes data from 32 subjects. The exact number of participants in the CMED may differ, but it usually encompasses a diverse group to ensure a wide array of facial expressions and emotions. These datasets cover a range of both basic and complex emotions. Common basic emotions captured include happiness, sadness, anger, fear, surprise, and disgust, which are universally understood and easily recognizable. Additionally, complex emotions like contempt, embarrassment, pride, and guilt are also recorded to offer a more thorough insight into human emotional expression.

CASME

The first dataset established by the Chinese Academy of Sciences (CAS) was CASME. It was recorded at 60 frames per second and includes 180 videos from 19 subjects. Thirteen females and twenty-two males, with an average age of 22.03 years ($\sigma = 1.60$) were enlisted for the study. The eight micro-expressions that have been recognized are: amusement, sadness, disgust, surprise, contempt, fear, repression, and tense. The experiments were limited to the top five classes (tense, disgust, happiness, surprise, and repression) with the greatest number of samples[64].

CFEE

This dataset contains 5,060 images of faces from 230 people which have been labeled with 15 complex emotions and 7 basic emotions.[65] There are 21 different emotion categories in the CFEE dataset. Compound emotions, like delightfully surprised and angry surprised, comprise multiple basic component categories. The Facial Action Coding System was utilized for the analysis of the obtained images. These 21 categories have distinct production processes that are nevertheless in line with the subordinate categories they represent[66].

FER-2013

This dataset was introduced at the 2013 International Conference on Machine Learning (ICML) by Pierre-Luc Carrier and Aaron Courville, the 2013 Facial Expression Recognition dataset (FER-2013) is available in the Kaggle dataset. The 35,887 grayscale 48x48-pixel images in the FER-2013 dataset remain in a spreadsheet with the pixel values of each image listed in a row of cells. After using Google to source images, they were categorized into various emotional classes including surprise, anger, disgust, fear, happiness, neutral, and sadness. Upon the conclusion of the challenge, 3,589 images designated for private testing were added to the dataset, which originally contained 28,709 images for training and 3,589 images for public testing. Public test images are employed for a range of purposes in published research projects, each using the FER-2013 dataset distribution for individual training, validation, or test sets[67,68].

3.2 Basic emotions

DEAP

The main emphasis of the DEAP dataset (Database for Emotion Analysis using Physiological Signals) lies in basic emotions. The DEAP dataset [69] has been split into two sections. The first contains the ratings from 120 one-minute music video excerpts that participants reviewed online using three criteria: arousal, valence, and dominance. The participants ranged in age from 14 to 16. The second accumulation of data comprises participant ratings, physiological recordings, and face films from an experiment in which 32 participants viewed a selection of the 40 music videos mentioned above. Every participant reviewed the videos as above, and physiological and EEG signals were recorded. Oval face videos were also captured for 22 subjects.

BP4D+

The main focus of the BP4D+ dataset is on basic emotions. A large-scale, multimodal emotion dataset is known as BP4D+. The FERA challenge of 2017 made use of it. There are 140 participants total, aged 18–66, including 58 male and 82 female participants. Eight physiological signals are present in total: heart rate, respiration (rate and voltage), blood pressure (diastolic, systolic, mean, and raw), and electrodermal (EDA). Ten target emotions including happiness, sadness, anger, disgust, embarrassment, astonished skepticism, fear, pain, and surprise are represented in the data for each subject [70].

3.3 Complex Emotion Datasets

RAF-DB

29,672 real-world images of faces that were extracted from Flickr are included in RAF-DB. 315 talented annotators have labeled each of the RAF-DB's images, with roughly 40 independent annotators labeling each image. The single-label subset and the multi-label subset are the two distinct subsets found in RAF-DB[71].

CEED

The CEED (Compound Emotion Expression Database) is mainly concentrated on capturing intricate emotions. Unlike datasets that center on fundamental emotions, CEED is tailored to document and examine emotional expressions that merge multiple basic emotions or encompass more subtle and socially influenced emotional states [72]. 480 images of eight young adult actors emulating nine complicated and six basic social-sexual emotional expressions are available in the complicated Emotion Expression Database (CEED). There is some racial variety among the actors, who are both male and female. Almost 800 individuals independently scored images to confirm how the expression was perceived [73].

These datasets are crucial for advancing research in emotion recognition systems, encompassing a variety of basic and complex emotional states. It is noteworthy that in some cases, datasets primarily designed for recognizing basic emotions could also be utilized effectively for understanding and analyzing complex emotional states, thereby expanding the applicability and scope of these datasets [47, 74]. We propose eight datasets corresponding to emotion recognition systems used in recent years, including facial expression, EEG and ECG signals, which are presented in Table 1.

Table 1. The summarization of the mentioned datasets is listed.

References	Dataset Name (complex/basic)	Emotions	Assessment Types Used
Zhao and Jiancheng[59]	CMED Complex and Basic emotion	Happiness, Disgust, Fear, Anger, Sadness, Surprise, Happily surprised, Sadly surprised, Fearfully surprised, Angrily surprised, Disgustedly surprised, Happily disgusted, Sadly fearful, Sadly angry, Sadly disgusted, Fearfully angry, Fearfully disgusted, Angrily disgusted	Not specified
Takalkar, Madhumita A., and Min Xu [75]	CASME Complex and Basic emotion	Contempt, Disgust, Fear, Happiness, Regression, Sadness, Surprise, Tense	Self-report ratings
[76]	CFEE Complex and Basic emotion	Angry, Fearful, Disgusted, Surprised, Happy, Sad and Neutral.	Facial Action Coding System
Han et al [77]	FER-2013 Complex and Basic emotion	Happy, Sad, Angry, Fear, Surprise, Disgust, and Neutral	Not specified

[78]	DEAP Basic	signal-based (EEG)	Self-report ratings, physiological recordings, EEG
Guerdelli et al [79]	BP4D+ Basic	happiness or amusement, surprise, sadness, startle or surprise, skeptical, embarrassment, fear or nervous, physical pain, angry and disgust)	Self-report ratings, physiological recordings, EEG
Yan et al [80]	RAF-DB Complex	neutral, happy, surprise, sad, anger, disgust, fear	Not specified
Benda, Margaret S., and K. S. Scherf. [73]	CEED Complex	six basic expressions (angry, disgusted, fearful, happy, sad, and surprised) and nine complex expressions (affectionate, attracted, betrayed, brokenhearted, contemptuous, desirous, flirtatious, jealous, and lovesick)	Self-report ratings

4. Preprocessing in CEMRS

In this section, the challenges of preprocessing in complex emotion recognition systems (CEMRS) will be discussed. This section is divided into two subsection: a) Physical Cues Preprocessing b) Physiological Cues Preprocessing

a) Physical Cues Preprocessing

This section focuses on the preprocessing difficulties related to physical cues, specifically facial expressions obtained via image processing. It delves into the intricacies of preparing facial images to derive significant features that represent various emotional states. Major challenges encompass managing fluctuations in lighting conditions, facial angles, and the range of facial expressions. Furthermore, it highlights the necessity for sophisticated techniques in feature extraction to effectively capture the subtleties of intricate emotions [81].

Initial stages involve detecting and aligning faces to ensure consistent positioning and orientation of facial features in images. Following face detection, normalization methods are utilized to counteract lighting variations by adjusting brightness, contrast, and color balance for standardization. Next, noise reduction techniques like Gaussian blurring or median filtering are applied to minimize unwanted image artifacts that could disrupt emotional cue recognition. Feature extraction is then carried out to capture important facial attributes crucial for emotion recognition, including identifying facial landmarks and extracting texture descriptors from facial regions. Given that feature vectors often consist of intricate data, dimensionality reduction techniques like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) are applied to decrease computational complexity and prevent overfitting. Through these preprocessing steps, facial expression recognition systems can enhance their accuracy, resilience, and ability to generalize in emotion classification tasks. This thorough preprocessing process ensures that the recognition system is well-prepared to provide consistent and precise results across various input images, enabling applications in affective computing, human-computer interaction, and physical research [82,83].

b) Physiological Cues Preprocessing

This section will address the difficulties linked with preprocessing physiological cues, such as those derived from electroencephalography (EEG) and electrocardiography (ECG) signals through signal processing. It will encompass the preprocessing procedures needed to cleanse and filter the physiological signals, extract relevant features, and alleviate artifacts to precisely capture the underlying emotional states. Challenges in this field may involve noise reduction, artifact elimination, feature extraction from intricate physiological signals, and ensuring the dependability and precision of the processed data for emotion recognition [84].

The preprocessing workflow involves several essential steps tailored to optimize the EEG and ECG signals for effective analysis. Firstly, artifact removal is imperative to eliminate noise and unwanted interference from the signals. EEG signals, for instance, are susceptible to various artifacts like eye blinks, muscle movements, and environmental electrical noise, while ECG signals can be affected by muscle activity and movement artifacts. Techniques such as independent component analysis (ICA) and adaptive filtering are commonly used to mitigate these artifacts, ensuring that the extracted signals accurately represent the underlying neural and cardiac activity. Following artifact removal, signal segmentation is performed to divide the continuous EEG and ECG recordings into shorter, temporally meaningful epochs, often aligned with specific stimuli or events that elicit emotional responses. This segmentation allows for the analysis of emotion-related signal patterns within defined time windows. After segmentation, feature extraction is conducted to capture relevant characteristics from the EEG and ECG signals indicative of different emotional states. For EEG signals, features may include spectral power, coherence, and asymmetry in specific frequency bands, while ECG features may encompass heart rate variability (HRV), amplitude variations, and timing intervals such as RR intervals. Dimensionality reduction techniques are then used to minimize the complexity of the feature space while keeping important information, such as principal component analysis (PCA) and wavelet processing. By implementing these preprocessing steps, EEG and ECG-based emotion recognition systems can significantly improve their accuracy, robustness, and generalization capabilities, allowing for the development of more reliable and versatile emotion classification models for use in affective computing, human-computer interaction, and psychological research [85,86,87].

5. Machine Learning and deep learning Models in Emotion Recognition: Limitations and Challenges

Most studies in emotion recognition have primarily focused on utilizing machine learning, including deep learning techniques, to recognize emotions based on basic features. However, emotions are inherently complex, and traditional machine learning models, including deep learning models, often struggle to capture this complexity effectively. While these models may perform reasonably well in recognizing basic emotions, they face several limitations when it comes to understanding and categorizing complex emotional states.

5.1 Limitations of Machine/Deep Learning Models

1. **Inability to Capture Complex Emotional States:** Traditional machine learning models, including deep learning models, often rely on predefined features and lack the flexibility to capture the nuances and subtleties of complex emotional states. As a result, they may struggle to distinguish between similar emotional expressions or interpret the context-dependent nature of complex emotions [88].
2. **Dependency on Handcrafted Features:** Many traditional machine learning and deep learning approaches require handcrafted features to be extracted from the data. These features may not fully represent the multidimensional nature of complex emotions, leading to a limited ability to generalize across different emotional contexts [89,90].
3. **Limited Adaptability to Individual Differences:** Traditional machine learning and deep learning models may not be adaptable enough to account for individual differences in expressing and experiencing emotions. They often rely on generic models that do not adequately capture the variability and subjectivity of emotional responses among different individuals [91].
4. **Overfitting and Generalization Issues:** Traditional machine learning and deep learning models may be prone to overfitting when trained on limited datasets, resulting in poor generalization to unseen data. This is particularly problematic in the case of complex emotions, where the variability and diversity of emotional expressions can be significant [92].

5.2 The Role of Meta-Learning

In the face of these challenges, meta-learning offers promising approaches to enhance the capabilities of complex emotion recognition systems. Meta-learning, which focuses on learning how to learn, can provide a framework for adapting traditional machine learning and deep learning models to better handle complex emotional states. By leveraging meta-knowledge acquired from a diverse range of emotional contexts, meta-learning algorithms can enhance the adaptability and generalization capabilities of emotion recognition systems. This approach enables

models to learn from previous experiences and rapidly adapt to new emotional contexts, thereby improving their performance in recognizing complex emotions [93,94].

Most studies have presented machine learning and deep learning techniques for recognizing emotion based on basic features but emotions are always complex. Few studies have worked on complex emotions based on basic features for recognition. In this scenario, another approach called meta-learning could play a vital role in these types of tasks. Fig 6. showcases our proposed taxonomy for emotion recognition systems.

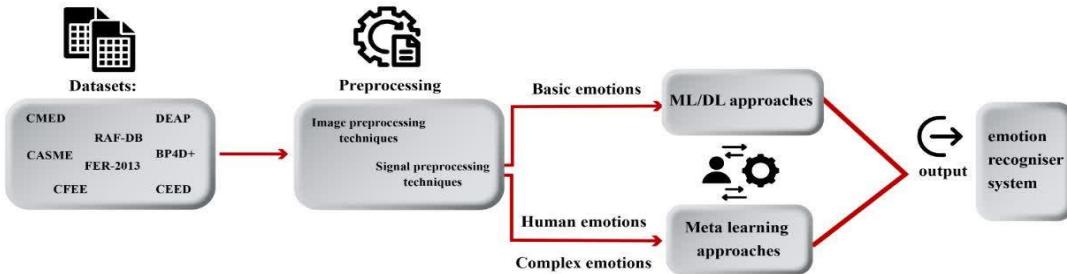


Figure 6. The taxonomy of emotion recognition systems.

5.3 Complex and basic emotion recognition system (methods and technologies)

The following sections are divided into basic and complex for introducing emotion recognizer systems based on machine learning, deep learning, and meta-learning methods.

5.4 Machine learning, Deep learning in Basic emotion Recognition

Machine learning approaches such as SVM, Random forest, KNN, Naive Bayes, etc in some studies utilized as feature selection or prediction tasks for emotion recognition[95]. Deep learning methods, such as recurrent neural networks (RNNs) and CNNs, have made significant strides in the field of computer vision in recent decades. These techniques based on deep learning have been used for problems related to recognition, classification, and feature extraction. By enabling "end-to-end" learning directly from input images, a CNN's primary benefit is to eliminate or greatly reduce reliance on physics-based models and/or other pre-processing approaches. Due to these factors, CNN has produced cutting-edge outcomes in several domains, such as emotion recognition tasks based on face expression, ECG, and EEG signals[96,97].

5.5 Meta-learning in complex emotion recognition

Meta-learning, commonly referred to as "learning to learn," is the process through which AI models acquire the ability to swiftly adjust to new environments or tasks while working with limited data. This capability is especially beneficial in the realm of complex emotion recognition, where challenges such as data scarcity and variability across different contexts frequently arise. Meta-learning comprises various methodologies: optimization-based, model-based, and metric-learning-based techniques [98].

Optimization-based techniques, including Model-Agnostic Meta-Learning (MAML), Reptile, and Almost No Inner Loop (ANIL), focus on identifying optimal initialization parameters that allow models to quickly converge on new tasks with minimal modifications. These techniques offer flexibility and adaptability across a variety of emotion recognition scenarios by effectively learning from a handful of examples [99,100].

Model-based techniques, such as recurrent and convolutional neural networks, integrate meta-learning principles directly into their architecture. These models are crafted to internally adjust to new tasks; however, they may encounter difficulties in generalizing to more complex or diverse emotional contexts due to their streamlined optimization processes.

Metric-learning-based techniques, like ProtoNet, RelationNet, and MatchingNet, depend on learning embedding functions that transform data into a space where classification can be performed using similarity metrics. These non-parametric approaches prove effective for emotion recognition as they facilitate rapid learning from sparse data by exploiting the relationships among data points [101].

In the subsequent sections, we will examine how meta-learning principles are implemented in Continual Learning, Few-shot Learning, Label Noise management, and Reinforcement Learning, all of which enhance the proficiency of models in recognizing intricate emotions:

Continual Learning guarantees that models can adapt to new emotional tasks over time without losing knowledge of previously learned tasks, which is essential in dynamic environments where emotions are subject to change.

Few-shot Learning confronts the obstacle of gaining insights from a small quantity of emotion-tagged information by empowering models to extrapolate proficiently from a tiny set of samples.

Label Noise management involves techniques for addressing noisy or inaccurately labeled emotional data, thereby improving the model's robustness and reliability in real-world emotion recognition applications.

Reinforcement Learning aids in optimizing emotion recognition policies through trial and error, allowing models to enhance their performance by engaging with their surroundings.

These methodologies underscore the importance of meta-learning in the advancement of complex emotion recognition. Through boosting the adaptability, resilience, and productivity of emotion recognition technologies, meta-learning is essential for confronting the intrinsic challenges that arise in this sector [98, 102]

5.6 Continual learning and Few-shot learning

Continual learning involves compiling research and methods to tackle the challenge of learning in scenarios where knowledge integration across endless streams of data needs to be considered, especially when the data distribution fluctuates over time [103,104]. In the realm of continual learning, models are crafted to acquire knowledge from new data while preserving previously gained knowledge, a task made difficult by catastrophic forgetting, a phenomenon where new learning can disrupt and overwrite previously absorbed information. To combat this issue, continual learning strategies often incorporate replay techniques, where a model is regularly trained on a combination of fresh data and selected samples of past data. These replay methods play a vital role in achieving the perfect equilibrium between stability—maintaining existing knowledge—and adaptability—effectively integrating new information. Furthermore, additional approaches like regularization techniques, which discourage alterations to crucial weights, and parameter isolation methods, which designate specific parts of the model for different tasks, are utilized to alleviate catastrophic forgetting and boost the performance of continual learning systems [105]. Fig 7. indicates the architecture of continual learning.

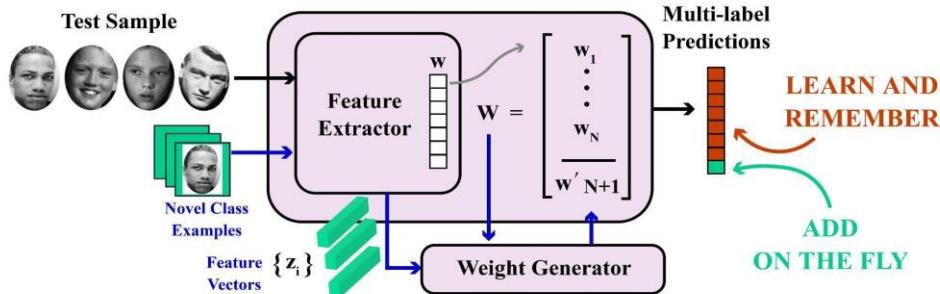


Figure 7. Continual learning in complex emotion recognition systems.

According to Fig 8, few-shot learning is the concept for developing an expanding algorithm from an insignificant sample set. Few-shot learning for facial emotion recognition has been established to decrease the intraclass distance and enhance the interclass distance[106].

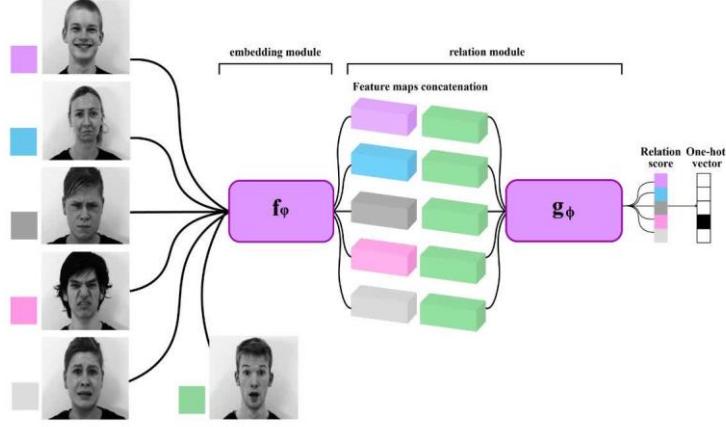


Figure 8. Few shot learning in complex emotion recognition systems.

5.7 Label noise

Label noise is a prevalent issue in real-world datasets caused by several factors, including the expense of the labeling process and the challenge of accurately classifying data. In the field of affective computing for complex emotion recognition, employing noisy labels addresses several challenges inherent in interpreting facial expressions. Concerning the subjective nature of human emotions, different annotators frequently provide diverse interpretations, leading to problems in labeling. Noisy labels help manage this ambiguity by allowing models to learn from a distribution of possible labels instead of a single, potentially erroneous one, thereby fostering more robust representations. Moreover, Fig 9 demonstrates that integrating label noise during training also enhances model robustness by leveraging techniques like label distribution learning, which improves generalization and mitigates the impact of incorrect labels. This approach is particularly beneficial for handling the complexity of real-world facial expressions, which often involve mixtures of basic emotions. Noisy labels mirror this variability and help models distinguish subtle emotional nuances. Additionally, methods like Face-Specific Label Distribution Learning (FSLDL) create augmented training samples with label distributions, broadening the range of expressions and viewpoints captured, thus enhancing the model's ability to generalize to new data. To prevent overfitting to noisy samples, techniques such as rank regularization and discriminative loss functions are employed, ensuring that the model focuses on more reliable samples and maintains overall performance. By addressing ambiguity and subjectivity, improving robustness, handling complex expressions, enhancing training with augmented data, and reducing overfitting, noisy labels significantly contribute to the advancement of complex emotion recognition systems in affective computing [107,108,109,110,111,112,113,114].

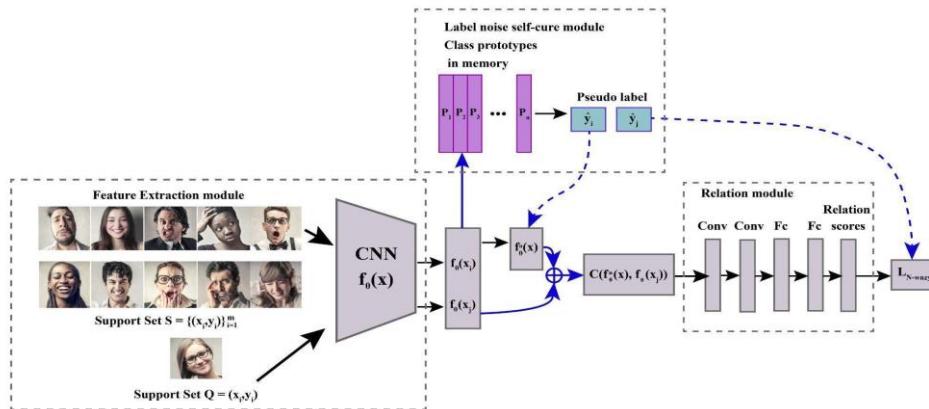


Figure 9. Architecture of meta-learning based on label noise.

5.8 Reinforcement learning

In affective computing for complex emotion recognition, employing reinforcement learning (RL) offers significant advantages by addressing key challenges. One major issue is handling unlabeled data, as complex emotions are often not explicitly labeled in datasets. RL, particularly through Deep Q-Networks (DQN), can learn from the environment via rewards and penalties, thus improving performance without the need for labeled data. Another challenge is identifying emotionally relevant intervals within data streams, such as video or physiological signals. RL-based segmentation dynamically learns to highlight these intervals, refining its strategy over time through rewards for accurate emotion recognition. It's also complex to integrate multimodal data, including facial expressions and physiological signals. RL excels in this by adaptively learning which signals are more indicative of specific emotions in varying contexts, enhancing the system's overall recognition capability. The system employs both facial expressions and physiological signals for emotion recognition. From facial expressions, it extracts confidence scores of seven basic emotions, valence-arousal (VA) levels, and ten action units (AUs). For physiological signals, it utilizes remote photoplethysmography (rPPG) to derive heart rate (HR) and heart rate variability (HRV) indices, along with EEG and ECG signals. This combination of facial and physiological data provides a comprehensive approach to recognizing complex emotions. The variability and subtlety of complex emotions often reduce recognition accuracy with traditional methods. RL optimizes segmentation and decision-making iteratively, focusing on the most informative data segments, thus improving recognition accuracy. Additionally, emotional states change rapidly, posing a challenge for static models. RL's adaptive learning allows for real-time updates in understanding and segmentation, making the system robust and effective in capturing dynamic emotional states in real-world scenarios. By addressing these challenges, RL significantly enhances the performance and reliability of complex emotion recognition systems in affective computing. Another significant advantage of using RL for complex emotion recognition is its ability to learn from sparse and unlabeled data. By focusing on key segments with significant emotional information, the RL-based segmentation module ensures that the decision module receives the most relevant data, leading to better recognition accuracy. This method is particularly effective for complex emotions, which are often subtle and not easily captured through simple observation. The RL approach allows the system to adapt and improve over time, refining its ability to detect and interpret complex emotional states. There is a step-by-step progression of emotions in interactions, which is comparable to how the action selected in reinforcement learning depends on the emotional state at that moment and the sequence of state transitions. The selected activity, representing the target's recognition results and the current emotional state, also impacts the reward function. This reinforcement learning module is essential for identifying the target emotion using characteristics of the appropriate emotion pair. By effectively managing unlabeled data, identifying key emotional intervals, integrating multimodal data, and improving recognition accuracy through iterative optimization, RL-based systems offer a comprehensive solution to the complexities of emotion recognition. The dynamic and adaptive nature of RL allows these systems to respond to rapid changes in emotional states, ensuring robust performance in real-world scenarios. The step-by-step progression and adaptation in RL mirror the evolving nature of human emotions, making it a powerful tool in affective computing. Reinforcement learning's ability to continuously improve through experience and feedback makes it particularly suited for the nuanced task of complex emotion recognition, providing a reliable and effective approach to understanding human emotions in various contexts which is highlighted in Fig 10.[115,116,117,118,119,120].

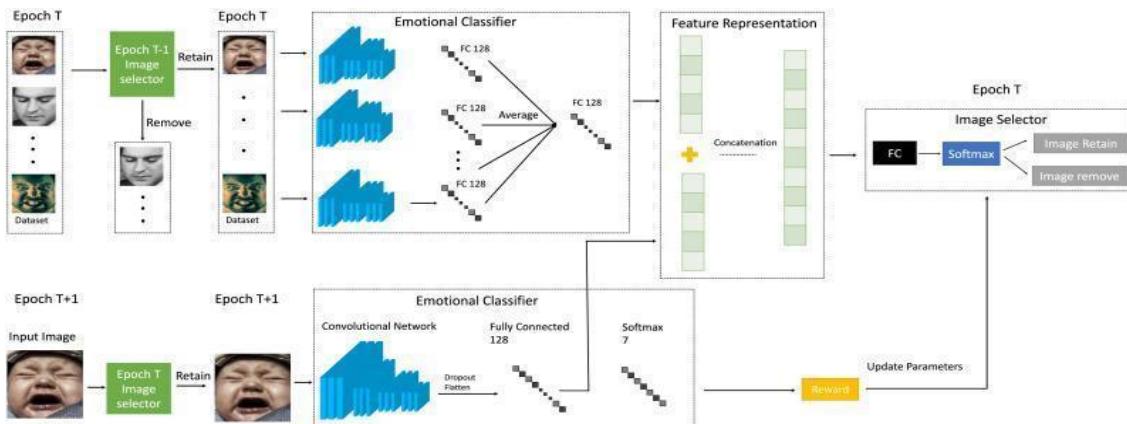


Figure 10. The application of reinforcement learning in complex emotion recognition systems.

6. A review of related works

The review of related works in basic and complex emotion recognition systems reveals a spectrum of advancements, from foundational models focusing on principal emotions. In this section, these provided studies underscore the evolution towards more accurate and context-aware systems.

6.1 Basic emotion recognition

In this section, we aim to review some studies related to AI approaches corresponding to the recognition of emotion by considering facial expressions, ECG, and EEG signals.

The construction of an artificial intelligence (AI) system that can recognize emotions from facial expressions. Jaiswal et al.[121] developed an AI system for emotion detection using deep learning, specifically a convolutional neural network (CNN) architecture. This model focuses on three main processes: face detection, feature extraction, and emotion classification. CNN's ability to automatically extract features and classify emotions showcases significant accuracy, achieving 98.65% on the JAFFE dataset and 70.14% on the FERC-2013 dataset. Key advantages of this model include high accuracy, the efficacy of deep learning in reducing manual feature extraction, and scalability. However, the model's limitations are its dependency on the quality and diversity of datasets, substantial computational requirements, and potential generalization issues across different populations and lighting conditions. Overall, while the CNN-based model marks a significant improvement over traditional methods, further enhancements are needed for broader application and real-world deployment.

In [122], a novel multi-modal input approach that uses color, depth, and thermal recording videos to estimate dimensional emotion states was described. Based on attention-boosted feature volumes, the proposed networks, dubbed multi-modal recurrent attention networks (MRAN), robustly distinguish facial expressions by learning spatiotemporal attention volumes. Utilizing the depth and heat sequences as guide priors, a selective focus on emotional discriminative regions on the color sequence was applied. Additionally, a brand-new multi-modal facial expression recognition benchmark called multi-modal arousal-valence facial expression recognition (MAVFER) comprised continuous arousal-valence scores matching to films of color, depth, and thermal recording. The outcomes of the experiments demonstrate that the approach was capable of producing state-of-the-art outcomes in color recording datasets for dimensional facial emotion recognition.

Zhang et al [123] studied the implementation of several deep learning models, such as long short-term memory (LSTM), convolutional neural networks (CNN), deep neural networks (DNN), and a hybrid model of CNN and LSTM (CNN-LSTM), to the subject of their research of EEG-based emotion recognition. The popular DEAP dataset was applied for evaluation. According to the experimental results, the CNN and CNN-LSTM models performed well in the categorization of EEG-based emotions, with accurate raw data extraction rates of 90.12 and 94.17%, respectively.

Bazgir et al [124] worked on the Deap dataset to build an AI system for the recognition of emotion based on electroencephalography (EEG) signals. Operating the discrete wavelet transform (DWT), EEG signals were divided into the gamma, beta, alpha, and theta frequency bands. Spectral features were then retrieved from each frequency band. To make the features mutually uncorrelated, principle component analysis (PCA) was used for the retrieved features while maintaining the same dimensionality as a transform. Emotional states were categorized by support vector machines (SVM), artificial neural networks (ANN), and k-nearest neighbors (KNN). With extracted features from ten EEG channels, the cross-validated SVM with radial basis function (RBF) kernel performs with 91.3% accuracy for arousal and 91.1% accuracy for valence in the beta frequency range.

The purpose of Sun and Lin's study [125] was to employ ECG signals to identify emotions. The data represented Four different emotions: happy, thrilling, tranquil, and tense. A finite impulse filter is then used to de-noise the raw data. To improve the accuracy of emotion recognition, The Discrete Cosine Transform (DCT) to extract characteristics from the collected data was applied. Support Vector Machine (SVM), Random Forest, and K-NN classifiers are investigated. The optimal parameters for the SVM classifier are found using the Particle Swarm Optimization (PSO) approach. The comparison of these classification techniques' findings shows that the SVM methodology recognizes emotions more accurately, which is useful in practical settings.

Nguyen et al [126] provided a new transfer learning approach utilizing PathNet to explore knowledge accumulation within a dataset and transfer insights from one emotion dataset to enhance overall performance to solve the generalization problem of developed deep learning models corresponding to a shortage of extensive

emotion datasets. The proposed system by passing different series of investigations on SAVEE and eINTERFACE datasets enhances emotion recognition performance according to experimental results, outperforming recent state-of-the-art methods that employ fine-tuning or pre-trained approaches. The highest recognition accuracy that the proposed system could obtain was 93.75% on SAVEE and 87.5% on eINTERFACE.

Albraikan et al [127] worked on the E4, and MAHNOB datasets to boost the classifier's accuracy rate by utilizing peripheral physiological signals. A hybrid sensor fusion method based on a stacking model was presented which enabled the simultaneous embedding of data from several sensors and emotion models within a model that was independent of the user. As a fundamental model for classifying emotions, WMD-DTW, a weighted multidimensional DTW, and the k-nearest neighbors algorithm were employed. On top of the two base models, a high-level classifier was learned using the ensemble methods. Applying a meta-learning methodology, were able to demonstrate the ensemble method performs more effectively than any particular method. The result showed recognizing valence and arousal emotions achieved 94.0% and 93.6% employing the MAHNOB dataset.

Table 2 provides an extensive overview of the techniques, results, advantages, and disadvantages of the featured study, along with additional pertinent research. Its aim is to assist in conducting a meticulous comparison, in order to clarify the pros and cons associated with each study. In this way, it aims to enhance comprehension of the present status of research in the domain of basic emotion recognition using physical and physiological cues.

Table 2. Overview of mentioned Studies and another studies on Basic Emotion Recognition Using Physical and Physiological Cues

Study	Methodology	Datasets	Metrics/Results	Advantages	Limitations
Jaiswal et al.	CNN for facial expressions	JAFFE, FERC-2013	70.14% (FERC-2013), 98.65% (JAFFE)	High accuracy on JAFFE	Lower accuracy on FERC-2013
Sun et al.	MRAN for multi-modal input	MAVFER	State-of-the-art results	Robust to multi-modal data	Computationally intensive
Zhang et al.	LSTM, CNN, DNN, CNN-LSTM	DEAP	90.12% (CNN), 94.17% (CNN-LSTM)	High classification accuracy	Potential overfitting
Bazgir et al.	DWT, PCA, SVM, ANN, KNN	DEAP	91.3% (arousal), 91.1% (valence)	Effective feature extraction	Dependence on frequency bands
Sun and Lin	DCT, SVM, Random Forest	Four emotions (ECG signals)	Highest accuracy with SVM	High practical accuracy	Limited emotion set
Nguyen et al.	PathNet for transfer learning	SAVEE, eINTERFACE	93.75% (SAVEE), 87.5% (eINTERFACE)	Improved generalization	Complexity in implementation
Albraikan et al.	Hybrid sensor fusion	E4, MAHNOB	94.0% (valence), 93.6% (arousal)	Effective multi-sensor integration	High computational cost

Lopes et al. [128]	CNN for Facial Expression Recognition	Custom dataset	Accuracy: 94.3%	High accuracy, robust model	Limited to specific dataset
Tripathi et al. [129]	DNN, CNN for EEG Emotion Recognition	DEAP	Accuracy: 86.5%	Utilizes multiple neural networks	Limited to DEAP dataset

6.2 Complex emotion recognition

In this section, an overview of selected papers focusing on meta-learning approaches for the recognition of complex emotions based on basic emotions will be presented. This review is prompted by the limited existing research on complex emotions. The selected papers are categorized into three distinct groups: Continual learning and few-shot learning will be discussed in part 1. Part 2 will cover label noises, while the investigation of reinforcement learning will be undertaken in the final part.

Part 1) Continual learning and few-shot learning

Angus and Nakisa [47] presented a new method based on continual learning and few-shot learning that improves and maintains its understanding of basic expression classes to recognize new compound expression classes accurately with a few training samples. Data augmentation, knowledge extraction, and a revolutionary Predictive Sorting Memory Replay to prevent catastrophic forgetting and enhance performance with fewer training samples were used. A considerable association between the activations of features in basic expressions and those in compound expressions was discovered by comparing the Grad-CAM heat maps of images of basic expressions with those of compound expressions. Continual learning outperforms non-continual learning methods in complicated face expression recognition, outperforming non-continual learning methods' state-of-the-art by 13.95%. The overall accuracy in new classes with 74.28% had demonstrated that continual learning for complex facial expression recognition played an essential role. This study is motivated by human cognition and learning patterns and it is the first to use few-shot learning for complex facial expression identification, attaining the state-of-the-art with 100% accuracy while only requiring one training sample for each expression class.

To showcasing significant improvements and practical applications in the domain of complex emotion recognition systems, Bhosale et al. [130] proposed a few-shot adaptation method from Electroencephalography (EEG) signals to address the lengthy calibration phase required by traditional Brain-Computer Interfaces (BCIs), hindering an optimal plug-and-play experience. Their model employed meta-learning to generalize well to new individuals with limited data, utilizing a few-shot learning framework trained on a small number of samples from previously unseen subjects, hence avoiding the need for retraining. Tested on the DEAP database with EEG recordings from 32 subjects watching music videos followed by emotion ratings, their method significantly improved emotion classification accuracy in terms of valence and arousal using only 20 reference samples from new subjects. Key contributions included quantifying the minimum calibration samples needed and introducing a 3-D convolutional recurrent embedding model to capture temporal relationships from spatially convolved EEG features. They explored various sampling strategies for support sets, finding that a combination of subject-dependent and subject-independent samples yielded competitive performance. In zero calibration scenarios, the model trained with subject-independent samples outperformed the supervised baseline. This system reduced the calibration burden and enhanced classification accuracy, advancing cross-subject EEG emotion recognition models and paving the way for more user-friendly and effective BCI applications.

By improving accuracy and robustness in recognizing micro-expressions for complex emotion recognition systems, a dual-branch meta-auxiliary learning method called LightmanNet for micro-expression recognition (MER) to address the challenges of limited data, subtle features, and individual differences in emotion detection was conducted by Wang et al [102]. LightmanNet utilizes a bi-level optimization process: in the first level, it learns task-specific MER knowledge through two branches. The primary branch focuses on learning MER features via primary MER tasks, while the auxiliary branch guides the model by aligning micro-expressions with macro-expressions, leveraging their spatial and temporal similarities. This dual-branch approach ensures the model learns meaningful features and avoids noise. In the second level, the model refines the task-specific knowledge, enhancing its generalization and efficiency. The method allows for quick acquisition of discriminative and generalizable MER knowledge from limited data. Extensive experiments demonstrated that LightmanNet significantly outperformed traditional, deep-learning, and meta-learning-based MER methods. The key contributions include addressing the data-level, feature-level, and decision-making-level challenges in MER,

proposing a novel dual-branch meta-auxiliary learning method to improve model generalization and efficiency, and demonstrating its superior robustness and effectiveness. This work advances the field of complex emotion recognition systems by improving accuracy and robustness in recognizing micro-expressions.

Part 2) Label Noise

In [48], Self-cure relation networks (SCRNet), a metric-based few-shot model that is resistant to label noise and capable of classifying facial images of new classes of emotions by only a few examples from each, was introduced as a solution to the complex emotion recognition via facial expressions problem that is demonstrated in a few-shot learning problem. By generating relation scores between the query image and the sparse samples of each new class, SCRNet establishes a distance metric based on deep information abstracted by convolutional neural networks and predicts an emotion category for a query image. Six basic emotion categories such as Happiness, Surprise, Sadness, Fear, Disgust, and Anger for facial expression detection to more intricate and compound emotions had developed. Given the difficulty in obtaining large datasets and the high level of expertise required for sophisticated facial expression interpretation In order to solve the label noise issue, SCRNet uses a class prototype maintained in external memory during the Meta training phase to assign corrected labels to noisy data. The effectiveness of the proposed approach has been verified on both synthetic noise datasets as well as public datasets.

Part 3) Reinforcement learning

A bionic two-system architecture for recognizing complicated emotions was proposed by Wu et al [32]. The design resembles how the human brain responds to challenges and makes decisions. A quick compound sensing module is System I. System II is a slower cognitive decision-making component that interacts with data processing more. System I has two branches: one for physiological measurement, which is a practical image-only implementation, and one for facial expression feature representation, comprising fundamental emotion, action units, and valence arousal detection. In System II, a decision module with segmentation is used to verify that the chosen time includes the occurrence of the emotion and to iteratively optimize the emotion information in a particular segment via reinforcement learning. By achieving an accuracy of 94.15% for basic emotion recognition on the BP4D database with five classes and an accuracy of 68.75% for binary valence arousal classification on the DEAP, the recommended approach outperforms advanced emotion recognition tasks. The recognition accuracy on both databases exceeds 70% for a selection of complicated emotions, which is a massive improvement.

7. Discussion

In this study, we reviewed research papers to provide an overview of current research on meta-learning approaches, focusing on three types of unstructured data including facial expressions, EEG, and ECG signals. Considering recent developments, the domain of complex emotion systems is still in its early stages, with very few papers examining meta-learning applications in these three domains. In complex emotion recognition systems, facial expressions, along with EEG and ECG signals, play a vital role. The preprocessing steps for facial expression data are similar to signal processing techniques but are tailored for image data. Standardizing input size and removing irrelevant background information through resizing and cropping facial images help expedite processing and reduce the computational burden. Additionally, normalization techniques ensure consistent pixel intensities across images, improving model performance by mitigating potential biases. Data augmentation methods like rotation, flipping, and adding noise to facial images enhance dataset diversity and prevent overfitting, similar to signal processing counterparts. Implementing these preprocessing steps on facial expression data helps extract more meaningful features, leading to accurate and robust emotion classification. Employing facial expression data, EEG, and ECG signals in a complex emotion recognition system delivers a holistic view of an individual's emotional state. This strategy utilizes supplementary data from various modalities to improve overall performance and dependability in real-world scenarios.

7.1 Evaluation Metrics

The development of CERS necessitates a robust evaluation framework tailored to the unique challenges of emotion recognition, where all proposed systems must experience evaluation on standardized datasets, comparing predicted emotion ratings or labels with ground truth. In classification problems [96,131], classifiers are frequently evaluated using a confusion matrix-based technique, as shown in Fig 11 .

individuals and society as a whole. It is essential to ensure that AI systems can offer understandable justifications for their decisions, reduce biases, and adhere to principles of fairness and equality to build trust and accountability. Various directions are being explored by researchers, such as explainable AI methods, fairness-aware learning algorithms, and cross-disciplinary partnerships with ethicists and social scientists, to tackle these crucial issues and encourage the responsible advancement and implementation of AI technologies. Efficiency and sustainability are crucial factors to consider in the development of intricate recognition systems. As models become more complex and computationally intensive, the focus on energy consumption and environmental impact grows. Researchers are actively seeking ways to improve model designs, create energy-efficient algorithms, and utilize hardware acceleration methods to boost the effectiveness and sustainability of AI systems. By prioritizing these initiatives, the industry can pave the path for the widespread integration of AI technologies while reducing their carbon footprint and resource usage. Finally, it is essential to carefully evaluate the broader societal impacts, such as privacy, security, and human well-being, during the creation and implementation of sophisticated recognition systems. Researchers are collaborating closely with various stakeholders to establish ethical guidelines, encourage inclusive and diverse data-gathering methods, and ensure transparent and responsible management of AI technologies. By proactively tackling these societal issues and incorporating ethical considerations into the core of AI research and advancement, the community can nurture a fairer, more accountable, and enduring future for AI technologies.

8.1 Conclusions and future research directions

In this study, we have delved into the effectiveness of the CERS in identifying human emotions within intricate tasks. Our research has showcased the success of CERS through different case studies, emphasizing its wide range of potential applications. This study marks a pioneering endeavor in this field, with no prior similar studies documented thus far. While previous research has focused on basic emotion recognition systems centered on facial expressions or physiological signals like EEG and ECG, our work goes beyond these methods to capture emotions within complex tasks. The case studies presented highlight the adaptability of CERS, suggesting its potential use in various sectors such as healthcare, education, and human-computer interaction. Our results indicate that even with partial implementation, CERS shows promising abilities in identifying underlying emotions linked to complex tasks. By combining data from multiple sensors, CERS can accurately deduce specific emotions, thereby improving human-computer interaction in intelligent systems. This study sets the foundation for future research aimed at improving the CERS using meta-learning techniques including incorporating advanced techniques such as few-shot learning and constant learning to allow for rapid adaptation and continuous development of emotion recognition across varied situations. Machine learning model optimization, particularly deep learning and ensemble approaches, has the potential to improve emotion recognition accuracy and robustness. Focusing on real-time emotion identification capabilities inside CERS, employing meta-learning approaches for quick adaptation to changing settings, and individualized emotion recognition models tailored to specific users' preferences are also critical. Furthermore, taking into account the ethical and societal implications of CERS deployment and defining responsible rules for its usage, including insights from meta-learning methodologies to assure fairness and transparency, is critical for the technology's ethical progress.

References

- [1] T. Fan, S. Qiu, Z. Wang, H. Zhao, J. Jiang, Y. Wang, J. Xu, T. Sun, N. Jiang, "A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition," *Computers in Biology and Medicine*, vol. 159, pp. 106938, 2023.
- [2] B. Gandhi, S. Saxena, P. Jain, "Emotion Recognition: A Review," in *Microelectronics, Circuits and Systems: Select Proceedings of Micro2021*, pp. 371-379, 2023.
- [3] P. Ekman, "An argument for basic emotions," *Cognitive & Emotion*, vol. 6, pp. 169-200, 1992.
- [4] P. Ekman and D. Cordaro, "What is meant by calling emotions basic," *Emotion Review*, vol. 3, no. 4, pp. 364-370, 2011.
- [5] C. E. Izard, "Basic emotions, natural kinds, emotion schemas, and a new paradigm," *Perspectives on Psychological Science*, vol. 2, no. 3, pp. 260-280, 2007.
- [6] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information*, vol. 44, no. 4, pp. 695-729, 2005.

[7] F. E. Oğuz, A. Alkan, T. Schöler, "Emotion detection from ECG signals with different learning algorithms and automated feature engineering," *Signal, Image and Video Processing*, pp. 1-9, 2023.

[8] T. Nokelainen, P. A. Airola, M. S. I. Elnaggar, "Physiological signal-based emotion recognition from wearable devices," *Health Technology*, 2023.

[9] B. Nakisa, M. N. Rastgoo, D. Tjondronegoro, and V. Chandran, "Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors," *Expert Systems with Applications*, vol. 93, pp. 143-155, 2018.

[10] B. Nakisa, "Emotion classification using advanced machine learning techniques applied to wearable physiological signals data," Ph.D. dissertation, Queensland University of Technology, Brisbane, Australia, 2019.

[11] M. N. Rastgoo, B. Nakisa, A. Rakotonirainy, F. Maire, and V. Chandran, "Driver stress levels detection system using hyperparameter optimization," *Journal of Intelligent Transportation Systems*, vol. 28, no. 4, pp. 443-458, 2024.

[12] M. N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, and V. Chandran, "Automatic driver stress level classification using multimodal deep learning," *Expert Systems with Applications*, vol. 138, 2019, Art. no. 112793.

[13] J. Kim, N. Bianchi-Berthouze, D. Patel, "Exploring User Experiences of Physical Activity Tracking Technology and the Implications for Design," *Human-Computer Interaction*, vol. 33, no. 3, pp. 267-304, 2018.

[14] Kranti Kamble, Joydeep Sengupta, "A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals," *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 27269-27304, 2023.

[15] B. Majhi, N. Das, and M. Chakraborty, "Analyzing emotional responses to audio-visual stimuli through heart rate variability analysis," in 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Feb. 2024, pp. 1-6.

[16] Z. S. Chen, I. R. Galatzer-Levy, B. Bigio, C. Nasca, Y. Zhang, "Modern views of machine learning for precision psychiatry," *Patterns*, vol. 3, no. 11, 2022.

[17] B. M. Booth, K. Mundnich, T. Feng, A. Nadarajan, T. H. Falk, J. L. Villatte, E. Ferrara, S. Narayanan, "Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: Framework for sensor selection, deployment, and management," *Journal of medical Internet research*, vol. 21, no. 8, p. e12832, 2019.

[18] S. Li, W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195-1215, 2020.

[19] Y. Huang, F. Chen, S. Lv, X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, 2019.

[20] N. Raut, "Facial emotion recognition using machine learning," 2018.

[21] R. W. Picard, *Affective Computing*, Cambridge, MA: MIT Press, 1997, p. 20.

[22] H. Wang, D. Nie, and B. L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94-106, 2014.

[23] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18-31, 2012.

[24] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on FER-2013," in *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications*, Cham: Springer, 2018, pp. 1-16.

- [25] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M. L. Shyu, S.-C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1-36, 2018.
- [26] A. Pramod, H. S. Naicker, and A. K. Tyagi, "Machine learning and deep learning: Open issues and future research directions for the next 10 years," in *Computational analysis and deep learning for medical care: Principles, methods, and applications*, 2021, pp. 463-490.
- [27] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition*, vol. 17, no. 2, pp. 484-495, 2008.
- [28] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18-37, Jan. 2010.
- [29] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 295-308, July-Sept. 2017.
- [30] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85-117, 2015.
- [31] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689-694, 2020.
- [32] Y.-C. Wu, L.-W. Chiu, C.-C. Lai, B.-F. Wu, and S. S. J. Lin, "Recognizing, Fast and Slow: Complex Emotion Recognition with Facial Expression Detection and Remote Physiological Measurement," *IEEE Transactions on Affective Computing*, 2023.
- [33] N. Ahmed, Z. Al Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, pp. 200171, 2023.
- [34] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Information Fusion*, vol. 102019, 2023.
- [35] Y. Ding, X. Tian, L. Yin, X. Chen, S. Liu, B. Yang, and W. Zheng, "Multi-scale relation network for few-shot learning based on meta-learning," in *International Conference on Computer Vision Systems*, Cham: Springer International Publishing, 2019, pp. 343-352.
- [36] X. He, J. Sygnowski, A. Galashov, A. A. Rusu, Y. W. Teh, and R. Pascanu, "Task agnostic continual learning via meta learning," *arXiv preprint arXiv:1906.05201*, 2019.
- [37] Z. Wang, G. Hu, and Q. Hu, "Training noise-robust deep neural networks via meta-learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4524-4533.
- [38] H. Li and H. Xu, "Deep reinforcement learning for robust emotional classification in facial expression recognition," *Knowledge-Based Systems*, vol. 204, p. 106172, 2020.
- [39] S. Gu, F. Wang, N. P. Patel, J. A. Bourgeois, and J. H. Huang, "A model for basic emotions using observations of behavior in *Drosophila*," *Frontiers in psychology*, vol. 10, p. 445286, 2019.
- [40] M. H. Black, N. T. Chen, O. V. Lipp, S. Bölte, and S. Girdler, "Complex facial emotion recognition and atypical gaze patterns in autistic adults," *Autism*, vol. 24, no. 1, pp. 258-262, 2020.
- [41] C. Zhu, P. Li, Z. Zhang, D. Liu, and W. Luo, "Characteristics of the regulation of the surprise emotion," *Scientific Reports*, vol. 9, no. 1, p. 7576, 2019.
- [42] A. Weatherall and J. S. Robles, "How emotions are made to do things," *How emotions are made in talk*, vol. 321, pp. 1-24, 2021.

[43] I. Chaidi and A. Drigas, "Autism, expression, and understanding of emotions: literature review," pp. 94-111, 2020.

[44] D. Keltner and B. N. Buswell, "Embarrassment: Its distinct form and appeasement functions," *Psychological Bulletin*, vol. 122, no. 3, pp. 250-270, 1997.

[45] B. Mesquita and N. H. Frijda, "Cultural variations in emotions: A review," *Psychological Bulletin*, vol. 112, no. 2, pp. 179-204, 1992.

[46] B. Mesquita, "Emotions in collectivist and individualist contexts," *Journal of Personality and Social Psychology*, vol. 80, no. 1, pp. 68-74, 2001.

[47] A. Maiden and B. Nakisa, "Complex facial expression recognition using deep knowledge distillation of basic features," arXiv preprint, arXiv:2308.06197, 2023.

[48] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*, New York, NY: Psychology Press, 2001.

[49] A. Wierzbicka, *Emotions across Languages and Cultures: Diversity and Universals*, Cambridge, U.K.: Cambridge University Press, 1999.

[50] A. T. Beall and J. L. Tracy, "Emotivational psychology: How distinct emotions facilitate fundamental motives," *Social and Personality Psychology Compass*, vol. 11, no. 2, p. e12303, 2017.

[51] A. Milone, L. Cerniglia, C. Cristofani, E. Inguaggiato, V. Levantini, G. Masi, M. Paciello, F. Simone, and P. Muratori, "Empathy in youths with conduct disorder and callous-unemotional traits," *Neural plasticity*, vol. 2019, no. 1, p. 9638973, 2019.

[52] R. Adolphs, "How should neuroscience study emotions? By distinguishing emotion states, concepts, and experiences," *Social cognitive and affective neuroscience*, vol. 12, no. 1, pp. 24-31, 2017.

[53] G. Šimić, M. Tkalčić, V. Vukić, D. Mulc, E. Španić, M. Šagud, F. E. Olucha-Bordonau, M. Vukšić, and P. R. Hof, "Understanding emotions: origins and roles of the amygdala," *Biomolecules*, vol. 11, no. 6, p. 823, 2021.

[54] N. B. Rothman and S. Melwani, "Feeling mixed, ambivalent, and in flux: The social functions of emotional complexity for leaders," *Academy of Management Review*, vol. 42, no. 2, pp. 259-282, 2017.

[55] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao et al., "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19-52, 2022.

[56] S. PS and G. Mahalakshmi, "Emotion models: a review," *International Journal of Control Theory and Applications*, vol. 10, no. 8, pp. 651-657, 2017.

[57] R. Vempati and L. D. Sharma, "A systematic review on automated human emotion recognition using electroencephalogram signals and artificial intelligence," *Results in Engineering*, p. 101027, 2023.

[58] M. S. Chaubey and N. Pathrotkar, "Facial Recognition Ai: A Powerful Tool For Emotion Detection And Characterization," *Journal of Data Acquisition and Processing*, vol. 38, no. 2, pp. 1914, 2023.

[59] Y. Zhao and J. Xu, "A convolutional neural network for compound micro-expression recognition," *Sensors*, vol. 19, no. 24, p. 5553, 2019.

[60] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013.

[61] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, Art. no. e86041, 2014.

[62] S.-T. Liong, Y.-S. Gan, C.-J. Wong, and K. Wong, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 70-75, Jan.-Mar. 2018.

- [63] A. K. Davison, C. Lansley, J. F. Cohn, H. Gunes, and B. Martinez, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116-129, Jan.-Mar. 2018.
- [64] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Analysis and Applications*, vol. 22, pp. 1331-1339, 2019.
- [65] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko et al., "Dominant and complementary emotion recognition from still images of faces," *IEEE Access*, vol. 6, pp. 26391-26403, 2018.
- [66] C.-L. Kim and B.-G. Kim, "Few-shot learning for facial expression recognition: a comprehensive survey," *Journal of Real-Time Image Processing*, vol. 20, no. 3, pp. 52, 2023.
- [67] G. P. Kusuma, J. Jonathan, and A. P. Lim, "Emotion recognition on fer-2013 face images using fine-tuned vgg-16," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 6, pp. 315-322, 2020.
- [68] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, and S. B. Musa, "The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi," in *2020 Fifth international conference on informatics and computing (ICIC)*, 2020, pp. 1-9.
- [69] S. Tripathi, S. Acharya, R. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on DEAP data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31, no. 2, pp. 4746-4752.
- [70] D. Fabiano and S. Canavan, "Emotion recognition using fused physiological signals," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 42-48.
- [71] A. Greco, N. Strisciuglio, M. Vento, and V. Vigilante, "Benchmarking deep networks for facial emotion recognition in the wild," *Multimedia tools and applications*, vol. 82, no. 8, pp. 11189-11220, 2023.
- [72] Du, S., Tao, Y., & Martinez, A. M. (2014). "Compound facial expressions of emotion." *Proceedings of the National Academy of Sciences*, 111(15), E1454-E1462.
- [73] M. S. Benda and K. S. Scherf, "The Complex Emotion Expression Database: A validated stimulus set of trained actors," *PloS one*, vol. 15, no. 2, p. e0228248, 2020.
- [74] X. Wang, Y. Wang, and D. Zhang, "Complex Emotion Recognition via Facial Expressions with Label Noises Self-Cure Relation Networks," *Computational Intelligence and Neuroscience*, vol. 2023, 2023.
- [75] M. A. Takalkar and M. Xu, "Image based facial micro-expression recognition using deep learning on small datasets," in *2017 international conference on digital image computing: techniques and applications (DICTA)*, 2017, pp. 1-7.
- [76] V. Mavani, S. Raman, and K. P. Miyapuram, "Facial expression recognition using visual saliency and deep learning," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 2783-2788.
- [77] B. Han, H. Kim, G. J. Kim, and J.-I. Hwang, "Masked FER-2013: Augmented Dataset for Facial Expression Recognition," in *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2023, pp. 747-748.
- [78] M. Khateeb, S. M. Anwar, and M. Alnowami, "Multi-domain feature fusion for emotion classification using DEAP dataset," *IEEE Access*, vol. 9, pp. 12134-12142, 2021.
- [79] H. Guerdelli, C. Ferrari, W. Barhoumi, H. Ghazouani, and S. Berretti, "Macro-and micro-expressions facial datasets: A survey," *Sensors*, vol. 22, no. 4, p. 1524, 2022.
- [80] H. Yan, Y. Gu, X. Zhang, Y. Wang, Y. Ji, and F. Ren, "Mitigating label-noise for facial expression recognition in the wild," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1-6.
- [81] P. Schulze, A.-K. Bestgen, R. K. Lech, L. Kuchinke, and B. Suchan, "Preprocessing of emotional visual information in the human piriform cortex," *Scientific Reports*, vol. 7, no. 1, p. 9191, 2017.

- [82] Samadiani, Najmeh, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. "A review on automatic facial expression recognition systems assisted by multimodal sensor data." *Sensors* 19, no. 8 (2019): 1863.
- [83] Li, Lixiang, Xiaohui Mu, Siying Li, and Haipeng Peng. "A review of face recognition technology." *IEEE access* 8 (2020): 139110-139120.
- [84] K. Yang, C. Wang, Y. Gu, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves, "Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1082-1097, 2021.
- [85] C. R. Rashmi and C. P. Shantala, "EEG artifacts detection and removal techniques for brain computer interface applications: a systematic review," *International Journal of Advanced Technology and Engineering Exploration*, vol. 9, no. 88, pp. 354-383, 2022.
- [86] A. Shoka, M. Dessouky, A. El-Sherbeny, and A. El-Sayed, "Literature review on EEG preprocessing, feature extraction, and classifications techniques," *Menoufia J. Electron. Eng. Res*, vol. 28, no. 1, pp. 292-299, 2019.
- [87] C. Dora and P. K. Biswal, "Engineering approaches for ECG artefact removal from EEG: a review," *International Journal of Biomedical Engineering and Technology*, vol. 32, no. 4, pp. 351-383, 2020.
- [88] A. Hassouneh, A. M. Mutawa, and M. Murugappan, "Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods," *Informatics in Medicine Unlocked*, vol. 20, p. 100372, 2020.
- [89] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301-1309, 2017.
- [90] U. Côté-Allard, E. Campbell, A. Phinyomark, F. Laviolette, B. Gosselin, and E. Scheme, "Interpreting deep learning features for myoelectric control: A comparison with handcrafted features," *Frontiers in bioengineering and biotechnology*, vol. 8, p. 158, 2020.
- [91] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical signal processing and control*, vol. 47, pp. 312-323, 2019.
- [92] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46-56, 2019.
- [93] F. Zhou, C. Cao, T. Zhong, and J. Geng, "Learning meta-knowledge for few-shot image emotion recognition," *Expert Systems with Applications*, vol. 168, p. 114274, 2021.
- [94] R. K. Gandhi, "Performance analysis of meta-learning and contrastive learning for speech emotion recognition."
- [95] A. Sepúlveda, F. Castillo, C. Palma, and M. Rodriguez-Fernandez, "Emotion recognition from ECG signals using wavelet scattering and machine learning," *Applied Sciences*, vol. 11, no. 11, p. 4945, 2021.
- [96] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, p. 401, 2018.
- [97] Md Rabiul Islam, Mohammad Ali Moni, Md Milon Islam, Md Rashed-Al-Mahfuz, Md Saiful Islam, Md Kamrul Hasan, Md Sabir Hossain, et al., "Emotion recognition from EEG signal focusing on deep learning and shallow learning techniques," *IEEE Access*, vol. 9, pp. 94601-94624, 2021.
- [98] W. Wang, J. Zhang, Z. Lin, L. Cui, and X. Zhang, "Meta-learning improves emotion recognition," in *Proceedings of the World Conference on Intelligent and 3-D Technologies (WCI3DT 2022)*, R. Kountchev, K. Nakamatsu, W. Wang, and R. Kountcheva, Eds. Singapore: Springer, 2023, vol. 323, pp. 123-135.

- [99] Y. Feng, J. Chen, J. Xie, T. Zhang, H. Lv, and T. Pan, "Meta-learning as a promising approach for few-shot cross-domain fault diagnosis: Algorithms, applications, and prospects," *Knowledge-Based Systems*, vol. 235, p. 107646, 2022.
- [100] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, vol. 2, no. 3, p. 4, 2018.
- [101] D. Nguyen, D. T. Nguyen, S. Sridharan, and others, "Meta-transfer learning for emotion recognition," *Neural Computing & Applications*, vol. 35, no. 13, pp. 10535–10549, 2023.
- [102] Jingyao Wang, Yunhan Tian, Yuxuan Yang, Xiaoxin Chen, Changwen Zheng, and Wenwen Qiang, "Meta-Auxiliary Learning for Micro-Expression Recognition," *arXiv preprint arXiv:2404.12024*, 2024.
- [103] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Information fusion*, vol. 58, pp. 52-68, 2020.
- [104] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). "Continual lifelong learning with neural networks: A review." *Neural Networks*, 113, 54-71.
- [105] Yifan Chang, Wenbo Li, Jian Peng, Bo Tang, Yu Kang, Ynjie Lei, Yuanmiao Gui, Qing Zhu, Yu Liu, and Haifeng Li, "Reviewing continual learning from the perspective of human-level intelligence," *arXiv preprint arXiv:2111.11964*, 2021.
- [106] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari, "A survey on few-shot class-incremental learning," *Neural Networks*, vol. 169, pp. 307-324, 2024.
- [107] Huan Yan, Yu Gu, Xiang Zhang, Yantong Wang, Yusheng Ji, and Fuji Ren, "Mitigating label-noise for facial expression recognition in the wild," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, IEEE, 2022.
- [108] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884-2896, 2018.
- [109] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.
- [110] Stanisław Saganowski, Bartosz Perz, Adam G. Polak, and Przemysław Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876-1897, 2022.
- [111] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6897-6906, 2020.
- [112] P. Sarkar and A. Etemad, "Self-supervised ECG representation learning for emotion recognition," in *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1541-1554, 2020.
- [113] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," in *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [114] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 222-237.
- [115] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1034-1047, 2021.
- [116] M. Zhao and Y. Zhang, "RL-Emotion: A Deep Reinforcement Learning Framework for Multimodal Emotion Recognition in Videos," in *IEEE Transactions on Affective Computing*, 2021.

[117] J. Chen, H. Wu, and S. Han, "Emotion Recognition Based on Eye Movement and EEG Using Deep Reinforcement Learning," in IEEE Access, vol. 8, pp. 165899-165910, 2020.

[118] X. Liu, J. Liu, and X. Luo, "Emotion Recognition Based on Facial Micro-Expressions Using Reinforcement Learning," in Journal of Visual Communication and Image Representation, vol. 64, p. 102621, 2019.

[119] Z. Wang and T. Zhang, "Emotion Recognition Using Deep Recurrent Neural Networks with RL-Based Feature Selection," in IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 3, pp. 668-680, 2018.

[120] Y. Zhang, Z. Zhang, and P. Li, "A Reinforcement Learning Approach to Multimodal Emotion Recognition Using Physiological Signals," in IEEE Transactions on Affective Computing, 2022.

[121] A. Jaiswal, A. K. Raju, and S. Deb, "Facial emotion detection using deep learning," in 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-5.

[122] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," in IEEE Transactions on Image Processing, vol. 29, pp. 6977-6991, 2020.

[123] Y. Zhang et al., "An Investigation of Deep Learning Models for EEG-Based Emotion Recognition," in Frontiers in Neuroscience, vol. 14, p. 622759, 2020.

[124] O. Bazgir, Z. Mohammadi, and S. A. H. Habibi, "Emotion recognition with machine learning using EEG signals," in 2018 25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME), 2018, pp. 1-5.

[125] B. Sun and Z. Lin, "Emotion recognition using machine learning and ECG signals," arXiv preprint arXiv:2203.08477, 2022.

[126] D. Nguyen, K. Nguyen, S. Sridharan, I. Abbasnejad, D. Dean, and C. Fookes, "Meta transfer learning for facial emotion recognition," in 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3543-3548.

[127] A. Albraikan, D. P. Tobón, and A. El Saddik, "Toward user-independent emotion recognition using physiological signals," in IEEE Sensors Journal, vol. 19, no. 19, pp. 8402-8412, 2018.

[128] A. T. Lopes, E. D. Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," in Pattern Recognition, vol. 61, pp. 610-628, 2017.

[129] S. Tripathi, S. Acharya, R. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on DEAP data," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 2, pp. 4746-4752, 2017.

[130] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Calibration Free Meta Learning Based Approach for Subject Independent EEG Emotion Recognition," in Biomedical Signal Processing and Control, vol. 72, p. 103289, 2022.

[131] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen, "EEG based emotion recognition: A tutorial and review," in ACM Computing Surveys, vol. 55, no. 4, pp. 1-57, 2022.

[132] F. Javier, E. Efren, J. Miguel, O. Roberto, and G. Manuel, "Evaluation of Machine Learning Algorithms for Classification of EEG Signals," in Technologies, vol. 10, no. 4, p. 79, 2022.

[133] P. Ackermann, C. Kohlschein, J. A. Bitsch, K. Wehrle, and S. Jeschke, "EEG-based automatic emotion recognition: Feature extraction, selection and classification methods," in 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), 2016, pp. 1-6.

[134] D. P. Russo, K. M. Zorn, A. M. Clark, H. Zhu, and S. Ekins, "Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction," in Molecular Pharmaceutics, vol. 15, no. 10, pp. 4361-4370, 2018.

Multi-scale spatiotemporal representation learning for EEG-based emotion recognition

Xin Zhou¹ and Xiaojiang Peng^{1,*}

Abstract—EEG-based emotion recognition holds significant potential in the field of brain-computer interfaces. A key challenge lies in extracting discriminative spatiotemporal features from electroencephalogram (EEG) signals. Existing studies often rely on domain-specific time-frequency features and analyze temporal dependencies and spatial characteristics separately, neglecting the interaction between local-global relationships and spatiotemporal dynamics. To address this, we propose a novel network called Multi-Scale Inverted Mamba (MS-iMamba), which consists of Multi-Scale Temporal Blocks (MSTB) and Temporal-Spatial Fusion Blocks (TSFB). Specifically, MSTBs are designed to capture both local details and global temporal dependencies across different scale subsequences. The TSFBs, implemented with an inverted Mamba structure, focus on the interaction between dynamic temporal dependencies and spatial characteristics. The primary advantage of MS-iMamba lies in its ability to leverage reconstructed multi-scale EEG sequences, exploiting the interaction between temporal and spatial features without the need for domain-specific time-frequency feature extraction. Experimental results on the DEAP, DREAMER, and SEED datasets demonstrate that MS-iMamba achieves classification accuracies of 94.86%, 94.94%, and 91.36%, respectively, using only four-channel EEG signals, outperforming state-of-the-art methods.

Index Terms—Electroencephalogram (EEG), emotion recognition, multi-scale, spatiotemporal feature.

I. INTRODUCTION

EMOTION recognition is pivotal for enhancing human-computer interaction and intelligent systems. Accurately identifying emotional states enables systems to respond more appropriately to human needs, thereby improving interaction naturalness and efficiency. EEG, as a non-invasive method for physiological signal acquisition, offers superior temporal resolution and continuity compared to other signals like facial expressions or voice, allowing real-time capture of human's emotional dynamics. In practical applications, EEG-based emotion recognition can facilitate mental health monitoring and diagnostic support by detecting abnormal patterns related to emotional disorders, providing objective indicators for clinical diagnosis [1].

Manuscript received: ***; Revised: ***; Accepted: ***.

This paper was recommended for publication by *** upon evaluation of the Associate Editor and Reviewers' comments.

This work is partially supported by the National Natural Science Foundation of China (62176165), the Stable Support Projects for Shenzhen Higher Education Institutions (20220718110918001), the Natural Science Foundation of Top Talent of SZTU (GDRC202131), the Basic and Applied Basic Research Project of Guangdong Province (2022B1515130009), and the Special subject on Agriculture and Social Development, Key Research and Development Plan in Guangzhou (2023B03J0172).

¹College of Big Data and Internet, Shenzhen Technology University, Shenzhen, 518118, China. (zhouxin, pengxiaojiang@sztu.edu.cn)

Extracting and analyzing discriminative spatiotemporal features from EEG signals is a challenging task due to the brain's complex spatial topology. Traditional approaches often involve manual extraction of domain-specific time-frequency features such as differential entropy (DE) [2], [3], power spectral density (PSD) [4], [5], and functional connectivity [6]. While these methods have advanced EEG emotion recognition, they are time-consuming, require extensive domain knowledge, and often lose valuable temporal information by compressing long time series into single single eigenvalue.

To address these limitations, deep learning methods have gained prominence for their end-to-end capabilities. For example, Cui et al. [7] utilized gated recurrent units combined with minimal class confusion for emotion recognition. Feng [8] and Li [9] integrated attention mechanisms into bidirectional long short-term memory (LSTM) modules to extract key temporal features from EEG sequences. Similarly, Du et al. [10] applied attention mechanisms with LSTM-generated feature vectors to automatically select appropriate EEG channels for emotion recognition. Other studies have framed physiological signal emotion recognition tasks as sequence-to-sequence multivariate time series prediction problems, employing advanced self-attention mechanisms to decompose signals into independent frequency and time-domain representations [11]. These approaches effectively capture useful temporal dependencies. Given that individual EEG time steps lack semantic meaning [12], the appropriateness of iterating or calculating mutual correlations among them is questionable. Inspired by this, we segment EEG signals into patches of different scales, aggregating time steps into subsequence-level patches to enhance local details and capture global relationships that single time points cannot provide.

Other deep learning methods have been utilized to construct spatial features from EEG signals, significantly enhancing emotion recognition accuracy. Typical spatial feature extraction methods include convolutional neural networks (CNNs) [13], [14]. For example, Li et al [15] employed a novel efficient convolutional block to reduce computational burden. Liu et al. [16] proposed a model named 3-D Convolutional Attention Neural Network (3DCANN), which consists of spatiotemporal feature extraction modules and EEG channel attention weight learning modules. This model effectively captures dynamic relationships between multi-channel EEG signals and the internal spatial relationships within these signals. Recent studies have shown that graph convolutions can effectively utilize brain topological structures for emotion recognition. Lin et al. [17] developed an improved graph convolution model combined with dynamic channel selection

to simulate information transmission in the brain. This model combines the advantages of one-dimensional convolution and graph convolution, capturing intra-channel and inter-channel EEG features and further modeling inter-regional brain connectivity by adding functional connectivity. Feng et al [8]. introduced a spatial graph convolution module that adaptively learns intrinsic connections between EEG channels using an adjacency matrix to extract spatial domain features. Additionally, researchers like Cui [18] and Deng et al. [19] explored the spatial information of adjacent and symmetrical channels from the perspective of whether EEG signals exhibit symmetrical emotional responses. These studies underscore the importance of understanding brain topological structures in EEG-based emotion recognition tasks.

The brain's complex structure results in EEG signals with time-varying spatial topology and temporal dependencies recorded through multiple electrodes. It is intuitive to use both temporal dependencies and spatial features as auxiliary information. For instance, a novel Attention-based Spatiotemporal Dual-Stream Fusion Network (ASTDF-Net) [20] has been employed to learn a joint latent subspace to capture the coupled spatiotemporal information in EEG signals. Cheng and Feng et al. proposed a hybrid model combining a Spatial-Graph Convolutional Network (SGCN) module and an attention-enhanced bidirectional Long Short-Term Memory (LSTM) module [8], and later designed a hybrid network comprising a Dynamic Graph Convolution (DGC) module and a Temporal Self-Attention Representation (TSAR) module, integrating spatial topology and temporal information [21]. Gong et al. [22] used a novel Attention-based Convolutional Transformer Neural Network (ACTNN), effectively integrating key spatial, spectral, and temporal information of EEG signals and cascading CNNs with transformers for emotion recognition tasks. Shen et al. [23] utilized multi-scale temporal self-attention modules to learn temporal continuity information while employing dynamic graph convolution modules to capture spatial functional relationships between different EEG electrodes. Although these integrated models consider both temporal and spatial features, they typically use two separate branches to extract spatiotemporal information, lacking adequate interaction between them.

Given these challenges, this article proposes a spatiotemporal fusion network called Multi-Scale Inverted Mamba (MS-iMamba), combining Multi-Scale Temporal Blocks (MSTB) and Temporal-Spatial Fusion Blocks (TSFB). The primary advantage of the proposed MS-iMamba is its ability to simultaneously leverage local details and global relationships in EEG signals without the need for complex statistical feature extraction, enhancing emotion recognition performance through adequate spatiotemporal dependency interactions. Specifically, MSTB divides EEG signals into multiple scale patches, using small-scale patches to capture fine local details and coarse global relationships, thereby utilizing complementary predictive capabilities in multi-scale observations. TSFB embeds the temporal dimension rather than the spatial dimension of reconstructed multi-scale EEG signals into a token and uses a Selective State-Space Model (SSM) to model spatiotemporal information. This mechanism fully integrates the spatiotem-

poral dependencies of both modules to enhance EEG emotion recognition.

The main contributions of this study are as follows:

- A plug-and-play MSTB is designed, which considers local details and global relationships without requiring traditional domain-specific time-frequency statistical feature extraction, providing a promising perspective for time dependency modeling in EEG emotion recognition.
- The proposed TSFB reflects on the modeling approach of the spatiotemporal characteristics of EEG signals, adequately considering the interaction between temporal dependencies and spatial features, offering a method that simultaneously integrates temporal and spatial features for EEG emotion recognition.
- Intra-subject and inter-subject experiments were conducted on three public datasets: DEAP, DREAMER, and SEED. The experimental results demonstrate that the proposed MS-iMamba outperforms various state-of-the-art methods using only four-channel EEG data.

The remaining sections of this article are organized as follows. Section II reviews the related work on multi-scale and spatiotemporal representation learning. Section III presents the pipeline of MS-iMamba. Section IV details the procedure of the conducted experiments and experimental results. A more in-depth discussion is provided in Section V. Finally, the study is concluded in Section VI.

II. RELATED WORK

A. Multi-Scale Representation Learning

Representing data in fine granularity has been widely adopted in time series prediction [12] and computer vision fields [24], [25]. EEG emotion recognition is essentially a time series prediction task, and considering an effective sequence representation approach is necessary. Nie et al. [12] argued that for time series data, single-point data lacks clear semantic information unlike words, making the computation of single-time-step correlations debatable. In natural language processing, it is also more effective to symbolize words in a sentence rather than each letter [26], [27]. This approach of aggregating single-point data into patches has been validated effectively in time series prediction tasks [28]. For instance, Wu et al. [29] addressed the limitations of 1-D time series by segmenting the sequence into short and long periods representations. These representations were embedded into the columns and rows of a 2-D tensor to capture inter- and intra-periodic variations, respectively, allowing easy modeling of 2-D variations using 2-D convolutional kernels. Chen et al. [30] highlighted the difficulty of capturing features across multiple scales when modeling time series with limited or fixed scales. Their proposed Pathformer model achieved multi-scale modeling by integrating time resolution and time distance, dividing the time series into different time resolutions and performing dual attention mechanisms at each scale to capture global correlations and local details as temporal dependencies.

Multi-scale representation learning has also been applied to EEG signal processing. Wang et al. [31] proposed the Multi-Scale Convolutional Neural Network-Dynamic Graph

Convolutional Network (AMCNN-DGCN) model to avoid the cumbersome manual feature extraction process. Jiang et al. [32] designed a novel Attention Mechanism-Based Multi-Scale Feature Fusion Network (AM-MSFFN) that considers high-level features at different scales to enhance the model's generalization capability across different subjects. To extract a comprehensive range of multi-class features from multi-channel EEG time series for accurate understanding of brain activity, Li et al. [33] introduced a Multi-Scale Attention Mechanism Fusion Convolutional Neural Network (MS-AMF), which extracts spatiotemporal multi-scale features from signals representing multiple brain regions and employs a dense fusion strategy to retain maximum information flow. These studies underscore the importance of multi-scale representation learning in EEG temporal modeling and demonstrate its potential in the field of emotion recognition.

B. Spatiotemporal Representation Learning

In the field of multivariate time series prediction, the fusion of spatiotemporal features has become a popular strategy for improving prediction accuracy. Numerous scholars have focused on effectively integrating temporal continuity with spatial correlations [34], [35]. Grigsby et al. proposed Space-timeformer [36], which transforms multivariate time series problems into a spatiotemporal sequence format. In this approach, each input token represents the value of a single variable at a given time step, allowing simultaneous learning of temporal and spatial relationships. Other works have modeled spatiotemporal relationships by transforming one-dimensional or multidimensional sequence data into tensors [37], [38]. Jin et al. [39] demonstrated that traditional methods, which process multichannel EEG signals into one-dimensional graphical features, limit the expressive capability of emotion recognition models. To address this issue, they introduced the G2G module, which transforms one-dimensional graphical data into two-dimensional grid data through numerical relationship encoding, using deep models like ResNet for subsequent tasks. Li et al. [40] employed dilated causal convolutional neural networks to extract nonlinear relationships between different time frames and used feature-level fusion to merge features from multiple channels, exploring potential complementary information between different views to enhance feature representation.

Recently, the integration of attention mechanisms and graph neural networks for EEG spatiotemporal modeling has gained increasing attention. Cheng and Feng have conducted extensive research in this direction. Initially, they proposed a model combining a Spatial Graph Convolution Network (SGCN) module with an attention-enhanced bidirectional Long Short-Term Memory (LSTM) module. This model's main advantage is its consideration of each brain region's biological topology, extracting representative spatiotemporal features from multiple EEG channels [8]. They subsequently designed a hybrid network comprising a Dynamic Graph Convolution (DGC) module and a Temporal Self-Attention Representation (TSAR) module, incorporating representative knowledge of spatial topology and temporal context into EEG emotion recognition

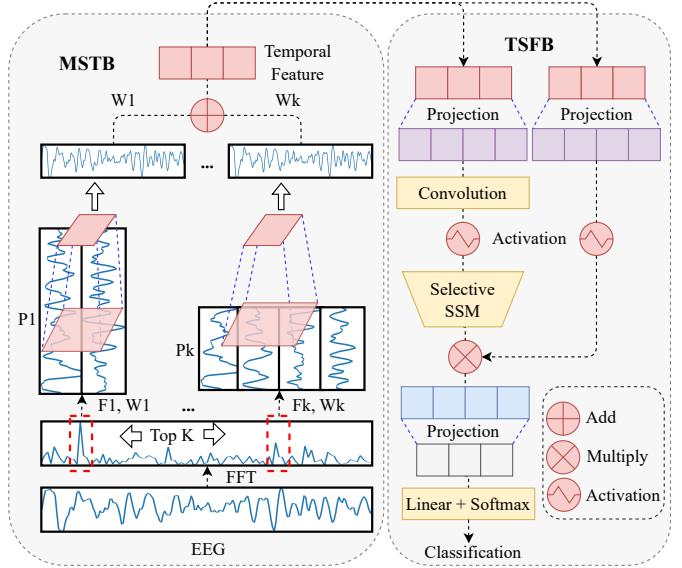


Fig. 1. Architecture of the MS-iMamba network for EEG emotion recognition. The network comprises two main modules: the Multi-Scale Temporal Block (MSTB) and the Temporal-Spatial Fusion Block (TSFB). The MSTB extracts multi-scale representations by converting the EEG signal into different frequency domain components and reshaping them into 2-D patches. These patches capture both local and global dependencies through convolution operations. The TSFB then integrates temporal and spatial information by embedding multiple time steps of the same channel into tokens, enabling effective feature extraction through the iMamba module, which combines a reversed embedding mechanism with a selective spatial state model (SSM).

tasks [21]. Recently, they equipped the Dense Graph Convolutional Network (DGC) with Joint Cross-Attention (JCA) for multimodal emotion recognition tasks, termed DG-JCA [41]. However, Zeng et al. [42] demonstrated that single-layer linear models unexpectedly outperformed complex Transformer-based models in time series prediction tasks. Liu et al. [43] reflected on this result, suggesting that for multivariate sequence data, points on different channels at the same time step record entirely different physical meanings or events, making embedding them into tokens inappropriate. They proposed the Inverted Transformer (iTransformer), which treats independent time series as tokens and captures multivariate correlations through self-attention to leverage spatiotemporal mutual information.

These methods share a common goal of revealing and utilizing the spatiotemporal features of time series data from different perspectives to achieve higher prediction accuracy. Each method has its specific application scenarios and advantages, but they all highlight the importance of spatiotemporal feature fusion in current research, providing a wealth of technical options and research directions for the field of EEG emotion recognition.

III. METHOD

In this section, we formalize the MS-iMamba network for EEG emotion recognition. As illustrated in Figure 1, the network consists of two primary modules: the Multi-Scale Temporal Block (MSTB) and the Temporal-Spatial Fusion Block (TSFB). Each module will be discussed in detail below.

A. Notations and Definitions

Let the EEG signals of each subject be represented by a 3-D matrix $S \in \mathbb{R}^{M \times T \times C}$, where M , T , and C denote the number of trials, sampling points, and channels, respectively. The matrix S is segmented into N samples of length L using a non-overlapping sliding window (thus, $T = N \times L$). The segmented EEG samples are denoted as $\mathcal{I} = \{(X_{ij}, Y_{ij}) \mid i = 1, 2, \dots, M; j = 1, 2, \dots, N\}$, where $X_{ij} \in \mathbb{R}^{L \times C}$ and $Y_{ij} \in \mathbb{R}$ represent the ground-truth label corresponding to X_{ij} . For the same trial, all N segments share the same label. Each segmented sample is denoted as $X_{1D} := X_{ij}$. The goal of EEG emotion recognition is to predict Y_{ij} given X_{1D} .

B. Multi-Scale Temporal Block (MSTB)

1) *Multi-Scale Representation*: Let X_{1D} denote an EEG signal of length L with C channels. Before representing this signal in a multi-scale format, we need to determine the patch sizes. To achieve this, we transform the original EEG signal into the frequency domain for analysis. Specifically, as shown in Equation 1:

$$\begin{aligned} A &= \mathcal{A}(FFT(X_{1D})), \\ \{f_1, f_2, \dots, f_k\} &= \text{argTop}_k(A), \\ p_i &= \lceil L/f_i \rceil, \quad i \in \{1, \dots, k\}, \end{aligned} \quad (1)$$

where FFT denotes the Fast Fourier Transform (FFT), and \mathcal{A} represents the amplitude calculation for each frequency. Since high-frequency regions often contain significant noise, we select only the top k frequencies with the highest amplitudes to avoid interference. The selected frequencies $\{f_1, f_2, \dots, f_k\}$ correspond to periods and amplitudes $\{p_1, p_2, \dots, p_k\}$ and $\{A_{f_1}, A_{f_2}, \dots, A_{f_k}\}$, respectively. The periods $\{p_1, p_2, \dots, p_k\}$ are used as the patch sizes for segmenting the EEG signal.

As illustrated in the left part of Figure 1, the original EEG signal is transformed into the frequency domain using FFT, with the red dashed boxes indicating the k frequencies with the highest amplitudes. We then calculate the weights for each frequency using Equation 2:

$$W_{f_i} = \{W_{f_1}, \dots, W_{f_k}\} = \text{Softmax}(A_{f_1}, \dots, A_{f_k}). \quad (2)$$

Next, the signal X_{1D} is segmented into patches of varying sizes and reshaped into a 2-D format, as shown in Equation 3:

$$X_{2D}^i = \text{Reshape}_{p_i, f_i}(\text{Padding}(X_{1D})), \quad i \in \{1, \dots, k\}, \quad (3)$$

where the padding operation ensures the original sequence can be divided into integer patches. The reshaped EEG signal is represented in a multi-scale format, $X_{2D}^i \in \mathbb{R}^{p_i \times f_i \times C}$, which denotes the i -th reshaped time series based on period p_i . The vertical and horizontal directions represent intra-patch and inter-patch variations, respectively. These variations capture local details and global relationships. Consequently, we obtain a set of 2-D tensors derived from different patches

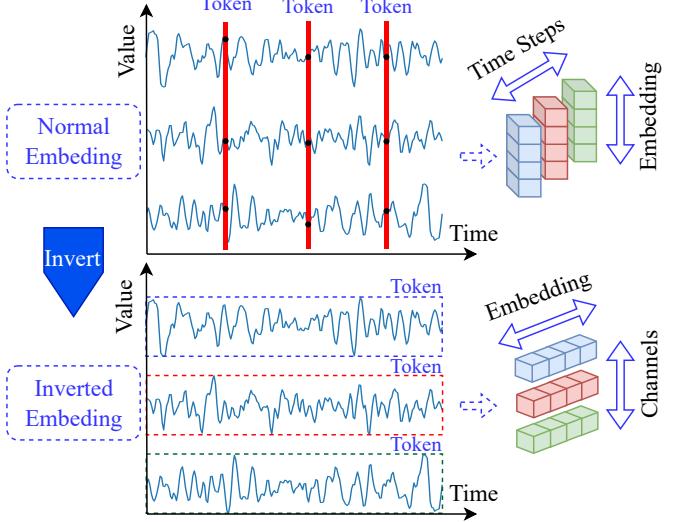


Fig. 2. Comparison between normal and inverted embedding mechanism. The top part illustrates the conventional embedding approach, where data from different channels at the same time step are mapped into a single token. The bottom part depicts the reversed embedding method, where multiple time steps of the same channel are mapped into a single token.

$\{X_{2D}^1, X_{2D}^2, \dots, X_{2D}^k\}$. This transformation facilitates capturing information at various distances, with larger p_i capturing longer temporal dependencies. Additionally, the reshaped tensors allow for efficient feature extraction using convolutional operations.

2) *Multi-Scale Perception*: The reshaped tensors are processed by the Multi-Scale Perception (\mathcal{MSP}) module, as shown in Equation 4:

$$X_{1D}^i = \text{Reshape}_{1, p_i \times f_i}(\mathcal{MSP}(X_{2D}^i)), \quad i \in \{1, \dots, k\}. \quad (4)$$

In this module, convolutional kernels of different sizes are employed. This mechanism allows the module to simultaneously perceive variations within the same patch and across patches with the same phase. After the convolution operations, we reshape the 2-D tensors back to the 1-D form X_{1D}^i . To assign different levels of attention to features extracted from patches corresponding to different frequencies, we perform a weighted sum of these multi-scale signals to obtain the final reconstructed multi-scale representation, as shown in Equation 5:

$$X_{1D} = \sum_{i=1}^k W_{f_i} \times X_{1D}^i. \quad (5)$$

This approach ensures that the features from various scales are effectively combined, enhancing the overall representation of the EEG signal for emotion recognition.

C. Temporal-Spatial Fusion Block (TSFB)

1) *Inverted Embedding Representation*: After obtaining the multi-scale representation of the EEG signal, we consider the interaction of temporal and spatial information. Generally, conventional methods embed data from different channels at

the same time step into a single token. As illustrated in the upper part of Figure 2, the conventional embedding method places points from different electrodes, each representing completely different events and physical meanings, into the same token. Specifically, at a certain time point, some channel data might be at a peak while others are at a trough. Embedding them into the same token not only fails to reveal valuable information due to the narrow focus of a single time point but also represents misaligned events as a single token.

Therefore, we adopt an inverted embedding method, as shown in the lower part of Figure 2. The inverted embedding method maps multiple time steps of the same channel into a single token. This event-driven representation not only considers information over longer time steps but also distinguishes data from different channels through separate tokens. The inverted embedding representation approach enhances the capacity to capture temporal dependencies and spatial relationships, ensuring a more comprehensive and meaningful interpretation of the EEG signals for emotion recognition.

This method is demonstrated through the following equations. Given a multi-scale EEG representation X_{1D} , we reshape it to consider the temporal and spatial interactions:

$$\hat{X}_{1D} = \text{Reshape}_{C,L}(X_{1D}). \quad (6)$$

Here, \hat{X}_{1D} represents that temporal steps and channels are reorganized to reflect the inverted embedding structure. To capture the dynamic interactions between temporal and spatial features, we apply a SSM to \hat{X}_{1D} . This inversion of the embedding representation and the fusion of temporal-spatial information using SSM enhance the ability to model the complex dependencies in EEG signals, leading to improved performance in emotion recognition tasks.

2) *iMamba*: After the inverted embedding operation, $X_{1D} \in \mathbb{R}^{L \times C}$ is transformed into $\hat{X}_{1D} \in \mathbb{R}^{C \times L}$. Next, we introduce the iMamba model, which consists of the inverted embedding mechanism and the SSM, specifically Mamba, to capture spatiotemporal correlations.

Mamba is inspired by continuous systems, mapping a 1-D sequence through a hidden state $h(t) \in \mathbb{R}^N$ to $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$. As shown in Equation 7, Mamba uses three parameter matrices $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathbf{B} \in \mathbb{R}^{d \times 1}$, and $\mathbf{C} \in \mathbb{R}^{1 \times d}$ (where d is the hidden dimension) to control this process. These parameters are analogous to the forget gate, input gate, and output gate mechanisms in LSTM. The parameter \mathbf{A} controls how much information is ignored, \mathbf{B} controls how the current input affects the hidden state, and \mathbf{C} controls the output flow of information:

$$\begin{aligned} H'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t). \end{aligned} \quad (7)$$

To adapt to discrete sequences, Mamba uses zero-order hold techniques, transforming \mathbf{A} and \mathbf{B} into their discrete versions via the time scale parameter Δ , as defined below:

$$\begin{aligned} \bar{\mathbf{A}} &= \exp(\Delta \mathbf{A}), \\ \bar{\mathbf{B}} &= (\Delta \mathbf{A})^{-1}(\exp(\Delta \mathbf{A}) - \mathbf{I}) \cdot \Delta \mathbf{B}. \end{aligned} \quad (8)$$

The discrete version is redefined as follows:

$$\begin{aligned} h_t &= \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t. \end{aligned} \quad (9)$$

For parallel processing, Mamba computes the output using the following convolution form:

$$\begin{aligned} \bar{\mathbf{K}} &= \left(\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \dots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}} \right), \\ \mathbf{y} &= \mathbf{x} * \bar{\mathbf{K}}, \end{aligned} \quad (10)$$

where L is the length of input sequence \mathbf{x} and \hat{K} is the structured convolution kernel.

Due to the inverted embedding operation, iMamba can extract both temporal and spatial features from the input data, fully considering spatiotemporal interactions. iMamba receives the input $\hat{X}_{1D} \in \mathbb{R}^{C \times L}$ and produces the output prediction \hat{Y}_i through the following calculation:

$$\hat{Y}_i = f(\text{iMamba}(\hat{X}_{1D})), \quad (11)$$

where $f(\cdot)$ is a linear classifier consisting of a Linear layer and a softmax operation. The cross-entropy loss is computed as follows:

$$\mathcal{L}_{cls} = - \sum_{i=1}^n \mathbb{1}_{[i=\hat{Y}_i]} \log(\hat{Y}_i), \quad (12)$$

where n denotes the number of categories and $\mathbb{1}_{[i=\hat{Y}_i]}$ equals 1 if the predicted class matches the true label, and 0 otherwise. Finally, the backpropagation algorithm is used to update the network parameters. The pseudocode of MS-iMamba is summarized in Algorithm 1.

IV. EXPERIMENT AND RESULTS ANALYSIS

A. Datasets

DEAP [44]: The DEAP dataset comprises multimodal data collected from 32 participants. Each participant watched 40 music videos while 32-channel EEG data and 8-channel peripheral physiological signals were recorded. Participants rated the videos on a scale from 1 to 9 for valence, arousal, dominance, and liking. Each video contains 60 seconds of data (excluding a 3-second baseline signal), which was downsampled to 128 Hz and filtered using a 4-45 Hz band-pass filter. We classified each metric into high and low categories using a threshold of 5. To augment the dataset, we segmented each signal into 1-second non-overlapping segments. In our experiments, we used only the frontal polar region channels FP1, FP2, AF3, and AF4.

DREAMER [45]: The DREAMER dataset also contains multimodal data from 23 participants. Each participant watched 18 video clips (ranging from 65s to 393s, with an average duration of 199s), while 14-channel EEG and 2-channel Electrocardiograph (ECG) signals were recorded. Participants rated valence, arousal, and dominance on a scale from 1 to 5. The signals were sampled at 128 Hz and filtered to 4-45 Hz using a band-pass filter. The EEG signals were then segmented into 1-second non-overlapping segments to

Algorithm 1 MS-iMamba for EEG Emotion Recognition

Require: EEG signal $S \in \mathbb{R}^{M \times T \times C}$
Ensure: Predicted label \hat{Y}

- 1: **Preprocessing:**
- 2: Slice S into non-overlapping windows to get samples $I = \{(X_{ij}, Y_{ij})\}$
- 3: **Multi-Scale Temporal Block (MSTB):**
- 4: Transform X_{ij} to frequency domain using FFT
- 5: $A = \mathcal{A}(\text{FFT}(X_{1D}))$
- 6: Select top k frequencies and their periods:
- 7: $\{f_1, f_2, \dots, f_k\} = \text{argTop}_k(A), p_i = \lceil L/f_i \rceil$
- 8: Calculate weights:
- 9: $W_{f_i} = \text{Softmax}(A_{f_1}, A_{f_2}, \dots, A_{f_k})$
- 10: Reshape X_{1D} into 2-D patches:
- 11: $X_{2D}^i = \text{Reshape}_{p_i, f_i}(\text{Padding}(X_{1D}))$
- 12: Apply multi-scale inception:
- 13: $X_{1D}^i = \text{Reshape}_{1, p_i \times f_i}(\mathcal{MSP}(X_{2D}^i))$
- 14: Combine multi-scale features:
- 15: $X_{1D} = \sum_{i=1}^k W_{f_i} \times X_{1D}^i$
- 16: **Temporal-Spatial Fusion Block (TSFB):**
- 17: Reverse embedding to reshape $\hat{X}_{1D} \in \mathbb{R}^{C \times L}$
- 18: **iMamba:**
- 19: Apply iMamba to capture spatio-temporal correlation:
- 20: $\hat{Y}_i = f(\text{iMamba}(\hat{X}_{1D}))$
- 21: Calculate cross-entropy loss:
- 22: $\mathcal{L}_{cls} = \sum_{i=1}^n \log(\hat{Y}_i) \cdot \mathbb{1}[i = \hat{Y}_i]$
- 23: Update network parameters using backpropagation.

expand the dataset. For DREAMER, we used four channels from the frontal polar and frontal regions: AF3, AF4, F7, and F8. Each metric was classified into high and low categories using a threshold of 3.

SEED [46]: The SEED dataset includes data from 15 participants, with 62-channel EEG data collected according to the international 10-20 system. Each participant conducted three sessions approximately one week apart, during which they watched 15 different film clips (each lasting about 4 minutes). These films elicited positive, neutral, and negative emotions as experimental stimuli. The data were downsampled to 200 Hz and filtered to 0-75 Hz, then segmented into 1-second non-overlapping segments. Only the frontal polar region channels FP1, FP2, AF3, and AF4 were selected for our experiments. Finally, to mitigate data drift across different channels, Z-score normalization was applied to all three datasets.

B. Training Protocol

In our experiments, we employed two different paradigms: intra-subject and inter-subject paradigm. For the intra-subject paradigm, we evaluated each participant's data individually, using 80% for training and 20% for testing. For the inter-subject paradigm, we combined and shuffled the data from all participants, splitting it into training and testing sets in a 4:1 ratio. Due to the SEED dataset comprising data from three different sessions, which significantly impacts experimental results, we also used intra-session and inter-session evaluation methods. Our training configuration included a batch size

of 32, the Adam optimizer with an initial learning rate of 1×10^{-3} , and 10 epochs of training. An adaptive learning rate strategy was employed to reduce the learning rate as the training loss decreased. Other hyperparameters, such as the number of network layers and Top-k, were set to 1 and 2, respectively. All experiments were conducted on an Intel Xeon Silver 4210R CPU @ 2.40GHz (x2) and an NVIDIA RTX A6000 GPU.

C. Baseline Model Selection

For our benchmark model selection, we chose several representative methods to compare against our model under the same experimental settings. These models are sourced from the Time Series Library (TSLib¹) and include the top three ranked models for classification tasks: TimesNet, Non-stationary Transformer (NTransformer), and Informer. Additionally, we included models characterized by linear structures and causal convolution structures, such as DLinear and TCN. Below is a brief introduction to these benchmark models:

- iTransformer [43]: iTransformer addresses the shortcomings of traditional Transformers in modeling spatiotemporal information by proposing an inverted Transformer structure that better considers spatiotemporal relationships.
- DLinear [42]: DLinear decomposes sequences into periodic and trend components, achieving impressive results in various time series tasks using a straightforward linear structure, outperforming many complex Transformer models and their variants.
- TimesNet [29]: TimesNet employs a multi-scale strategy to transform time series from 1-D to 2-D format, capturing both intra-period and inter-period variations.
- NTransformer [47]: This model designs non-stationary attention mechanisms to recover inherent non-stationary information in time dependencies through distinguishable attention learned from the raw sequences.
- Informer [48]: Informer utilizes sparse attention and a self-distillation mechanism to reduce the computational complexity of attention maps to logarithmic levels.
- TCN [49]: TCN introduces the concept of dilated causal convolutions, which are favored for expanding the receptive field without increasing computational burden.

D. Intra-subject Experiment Results

As shown in Table I, MS-iMamba demonstrates outstanding performance on both the DEAP and DREAMER datasets, significantly outperforming other models in most metrics. Specifically, it achieves the highest accuracy in DEAP (valence) at 94.69%, DEAP (arousal) at 95.03%, and DREAMER (valence) at 94.54%. It also achieves the second-highest accuracy in DREAMER (arousal) at 95.34%, underscoring its robustness and effectiveness in emotion recognition tasks. This makes MS-iMamba an excellent choice for applications requiring high-precision valence and arousal detection from the DEAP and DREAMER datasets. Notably, the linear model DLinear

¹TSLib: <https://github.com/thuml/Time-Series-Library>

TABLE I
PERFORMANCE COMPARISON OF MODELS ON DEAP AND DREAMER DATASETS (INTRA-SUBJECT)

Model Name	DEAP (valence)	DEAP (arousal)	DREAMER (valence)	DREAMER (arousal)
iTransformer	79.10%	81.35%	77.60%	79.78%
Dlinear	90.77%	91.89%	93.73%	95.47%
TimesNet	87.32%	88.05%	84.69%	88.47%
NTransformer	85.01%	87.01%	84.51%	86.25%
Informer	87.27%	88.39%	86.48%	89.47%
TCN	88.07%	89.24%	84.13%	88.63%
MS-iMamba	94.69%	95.03%	94.54%	95.34%

also performs well in this context, second only to MS-iMamba, and even achieving the highest accuracy in DREAMER (arousal). Surprisingly, TCN surpasses several Transformer-based models, while TimesNet performs comparably to them.

TABLE II
PERFORMANCE COMPARISON OF MODELS ON SEED DATASET (INTRA-SUBJECT, INTER-SESSION AND INTRA-SESSION)

Model Name	Inter session	Session 1	Session 2	Session 3
iTransformer	56.55%	60.18%	62.17%	54.28%
Dlinear	64.51%	85.80%	88.71%	81.07%
TimesNet	70.21%	79.79%	78.58%	68.93%
NTransformer	65.10%	75.20%	76.84%	63.15%
Informer	66.01%	79.15%	81.12%	68.84%
TCN	67.47%	75.36%	73.31%	60.33%
MS-iMamba	92.60%	93.24%	93.19%	87.67%

Table II presents the accuracy of different models across four sessions: inter-session, session 1, session 2, and session 3. MS-iMamba consistently maintains the highest accuracy in all sessions, demonstrating its strong performance and adaptability to various session conditions. The most significant improvement is observed in the inter-session scenario, where MS-iMamba outperforms the second-best model by approximately 22.39%. This substantial advantage highlights MS-iMamba's exceptional ability to generalize across different session data. In specific session scenarios, MS-iMamba surpasses the second-best model by 7.44%, 4.48%, and 6.60%, respectively. DLinear consistently ranks second in sessions 1, 2, and 3, indicating its reliability and effectiveness, although it lags noticeably in the inter-session scenario. TimesNet shows relatively high accuracy in the inter-session scenario (70.21%), but its performance declines in subsequent sessions, indicating potential limitations in session-specific contexts. Other models, such as iTransformer, NTransformer, and TCN, exhibit lower and more variable performance across sessions, indicating less consistency compared to MS-iMamba and DLinear. Overall, MS-iMamba demonstrates superior performance across all scenarios, significantly outperforming other models, particularly under inter-session conditions. This consistent and robust performance makes MS-iMamba an exceptional model for tasks requiring high accuracy under different session conditions. DLinear emerges as a strong contender, particularly effective in single-session scenarios, but falls short in terms of generalization compared to MS-iMamba.

E. Inter-subject Experiment Results

TABLE III
PERFORMANCE COMPARISON OF MODELS ON DEAP AND DREAMER DATASETS (INTER-SUBJECT)

Model Name	DEAP (valence)	DEAP (arousal)	DREAMER (valence)	DREAMER (arousal)
iTransformer	63.82%	65.65%	64.19%	74.29%
Dlinear	65.81%	66.28%	71.47%	78.21%
TimesNet	67.58%	69.17%	66.39%	75.65%
NTransformer	66.24%	68.76%	67.60%	75.75%
Informer	65.31%	65.98%	67.84%	74.82%
TCN	75.80%	78.87%	72.24%	79.01%
MS-iMamba	86.04%	85.94%	81.90%	87.04%

Table III presents the performance of various models under inter-subject conditions on the DEAP and DREAMER datasets. Compared to the intra-subject paradigm, the inter-subject setting poses a greater challenge for model generalization. MS-iMamba consistently outperforms other models in both intra-subject and inter-subject conditions, demonstrating significant robustness and generalization capability. However, due to increased data variability, all models exhibit a performance drop when transitioning from intra-subject to inter-subject conditions. Despite its excellent performance in intra-subject scenarios, DLinear shows a notable decline in inter-subject settings, highlighting potential limitations in handling data from different subjects. TCN maintains relatively stable performance across both conditions, making it a reliable choice, albeit not the top-performing one.

TABLE IV
PERFORMANCE COMPARISON OF MODELS ON SEED DATASET (INTER-SUBJECT, INTER-SESSION, AND INTRA-SESSION)

Model Name	Inter session	Session 1	Session 2	Session 3
iTransformer	38.76%	48.11%	47.80%	43.75%
Dlinear	39.82%	44.88%	44.50%	42.66%
TimesNet	47.39%	57.88%	55.11%	51.59%
NTransformer	43.68%	53.04%	48.79%	45.95%
Informer	43.46%	50.60%	47.61%	46.01%
TCN	68.25%	71.23%	50.35%	46.57%
MS-iMamba	86.10%	93.71%	94.54%	89.70%

We also conducted inter-subject experiments on the SEED dataset, along with inter-session and intra-session experiments, with results presented in Table IV. MS-iMamba outperforms the second-best model by 17.85%, 22.48%, 40.43%, and 38.11% in inter-session, session 1, session 2, and session 3, respectively. Although the accuracy of MS-iMamba in the inter-session experiment decreases by 6.5% compared to intra-subject conditions, its performance in intra-session experiments increases, while other models experience significant drops. Interestingly, while the DLinear model performs impressively in intra-subject experiments, it disappoints in inter-subject experiments, displaying the opposite pattern to TCN.

These results indicate that linear models are only suitable for scenarios with simple data structure distribution. Additionally, despite the multi-scale and inverted spatiotemporal structures used by TimesNet and iTransformer, their performance re-

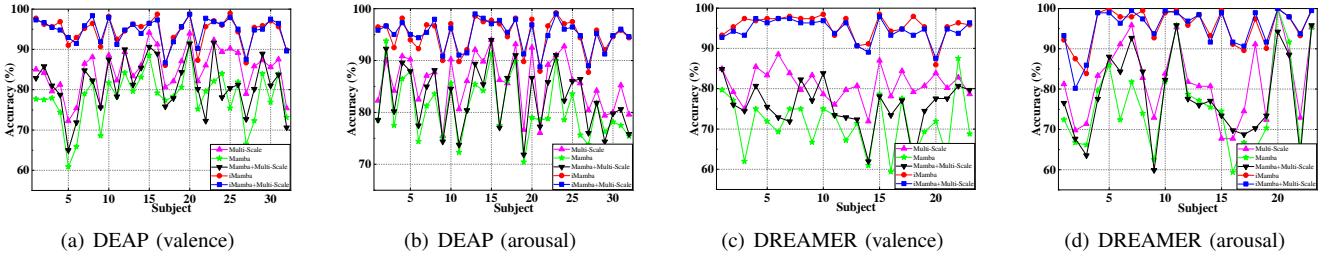


Fig. 3. Performance of different MS-iMamba variants on the DEAP and DREAMER datasets under intra-subject conditions.

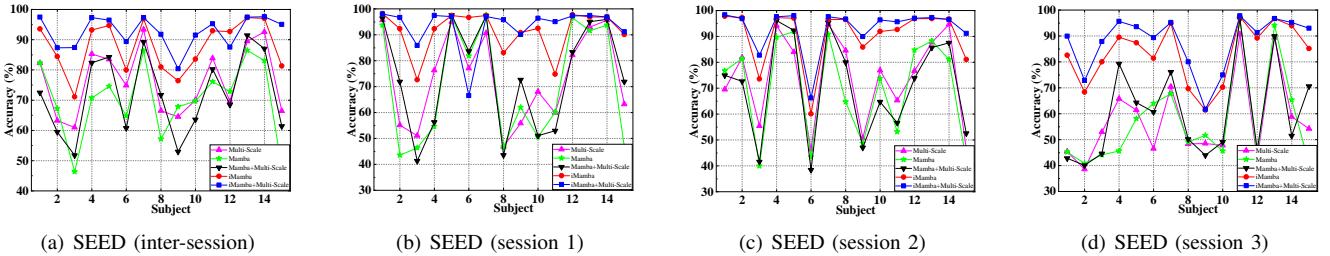


Fig. 4. Performance of different MS-iMamba variants on the SEED dataset across four session modes under intra-subject conditions.

mains unsatisfactory. Overall, MS-iMamba consistently outperforms other models across nearly all datasets in both intra-subject and inter-subject experiments, showcasing superior robustness and generalization capabilities.

TABLE V
CONFIGURATIONS OF MS-IMAMBA VARIANTS USED IN THE ABLATION EXPERIMENTS

Variants	MSTB	Mamba	Inverted Embedding
Multi-Scale	✓		
Mamba		✓	
Multi-Scale+Mamba	✓	✓	
iMamba		✓	✓
Multi-Scale+iMamba	✓	✓	✓

F. Ablation Study

To validate the effectiveness of each component in MS-iMamba, we conducted ablation experiments using five different configurations, as shown in Table V. We compared the performance of these five variants under both intra-subject and inter-subject conditions.

1) *Intra-subject Results:* We visualized the performance of the five variants across three datasets, with the results illustrated in Figures 3 and 4. Figures 3(a) and 3(b) show the accuracy of each individual's data on valence and arousal in the DEAP and DREAMER datasets, respectively. From the figures, we observe that the Mamba (green) performs the worst, while the Mamba with MSTB (black) shows slight improvement. However, both are outperformed by the variant using only MSTB (pink). This indicates that MSTB can effectively extract temporal features for emotion classification but does not integrate well with Mamba. The iMamba variant with inverted embedding (red) exhibits significant improvement, closely approaching the performance of MS-iMamba. These

results suggest that in intra-subject scenarios on the DEAP and DREAMER datasets, using inverted embedding to consider spatiotemporal interactions is more beneficial than using multi-scale features. Figure 4 displays the performance of these variants in four different session modes on the SEED dataset. The results indicate that Mamba shows significant improvement with the addition of MSTB and inverted embedding, with the latter providing a more substantial effect. These findings validate the effectiveness of MSTB and inverted embedding across all three datasets.

2) *Inter-subject Results:* We evaluated the performance of different MS-iMamba variants in inter-subject scenarios on the three datasets. Table VI shows that the combinations Mamba+Multi-Scale and iMamba+Multi-Scale, equipped with MSTB, achieve average accuracy improvements of 1.73% and 4.45%, respectively, compared to their counterparts without MSTB (Mamba and iMamba). The variants with inverted embedding (iMamba and iMamba+Multi-Scale) show average accuracy increases of 18.57% and 21.29%, respectively, compared to the Mamba and Mamba+Multi-Scale variants. The combined use of both mechanisms in iMamba+Multi-Scale (i.e., MS-iMamba) results in average accuracy improvements of 17.65% and 23.02% over the single-use Multi-Scale and Mamba variants. Overall, in inter-subject conditions, using MSTB, inverted embedding, or their combination leads to improved recognition performance on the DEAP, DREAMER, and SEED datasets.

G. Comparison with State-of-the-Art Methods

We compared MS-iMamba against state-of-the-art methods, and Table VII presents the average classification accuracies of various models on the DEAP, DREAMER, and SEED datasets. The input feature types include raw data (Raw), power spectral density (PSD), and differential entropy (DE). Among models utilizing feature extraction and all-channel

TABLE VI
PERFORMANCE OF DIFFERENT MS-iMAMBA VARIANTS ON THE DEAP, DREAMER, AND SEED DATASETS UNDER INTER-SUBJECT CONDITIONS.

Datasets	DEAP (valence)	DEAP (arousal)	DREAMER (valence)	DREAMER (arousal)	SEED (Inter session)	SEED (Session 1)	SEED (Session 2)	SEED (Session 3)	Mean
Multi-Scale	71.20%	72.29%	70.81%	77.84%	57.90%	75.87%	73.89%	63.92%	70.47%
Mamba	67.70%	71.28%	65.91%	74.76%	53.36%	67.25%	67.42%	53.10%	65.10%
Mamba+Multi-Scale	69.51%	71.71%	66.21%	74.62%	56.49%	66.54%	70.25%	59.33%	66.83%
iMamba	84.48%	85.51%	77.10%	81.21%	82.07%	91.05%	89.99%	77.98%	83.67%
iMamba+Multi-Scale	86.04%	85.94%	81.90%	87.04%	86.10%	93.71%	94.54%	89.70%	88.12%

TABLE VII
AVERAGE CLASSIFICATION ACCURACY COMPARED WITH THE
STATE-OF-THE-ART METHODS ON THREE DATASETS

Models	Input	Channels	DEAP	DREAMER	SEED
CRAM [15]	Raw	All	85.78%	92.65%	-
JO-CapsNet [50]	Raw	All	94.36%	-	-
DGCNN [3]	PSD	All	91.07%	85.39%	90.40%
IAG [51]	PSD	All	-	90.89%	-
V-IAG [52]	PSD	All	-	<u>92.96%</u>	-
EESCN [53]	DE	All	<u>94.81%</u>	-	-
TAE [54]	DE	30%	66.29%	-	-
ATDD-LSTM [10]	DE	All	-	-	91.08%
CSGNN [17]	DE	20%	83.39%	-	83.93%
Ours	Raw	4	94.86%	94.94%	91.36%

EEG data, EESCN [53], V-IAG [52], and ATDD-LSTM [10] achieved the highest accuracies on the three datasets, with 94.81%, 92.96%, and 91.08%, respectively. TAE [54], masking 70% of the data and using the remaining 30%, reached an accuracy of 66.29% on DEAP, while CSGNN [17], retaining only 20% of the channels, achieved accuracies of 83.39% on DEAP and 83.93% on SEED. These models performed poorly under conditions of incomplete data. In contrast, MS-iMamba, without using handcrafted features and relying on just four channels, achieved or exceeded the performance of these models. This demonstrates that our model effectively utilizes limited channel information to achieve high-precision classification. MS-iMamba consistently outperformed the state-of-the-art models across all datasets, highlighting the advanced nature and efficacy of our feature extraction and classification algorithms.

V. DISCUSSION

In this study, we designed MS-iMamba for EEG-based emotion recognition, incorporating two main components: Multi-Scale Temporal Blocks (MSTB) and Temporal-Spatial Feature Blocks (TSFB). MSTB and TSFB are utilized to capture multi-scale temporal features and spatiotemporal interactions, respectively. We replaced traditional manual time-frequency feature extraction with MSTB and introduced a novel approach to handle spatiotemporal information. The proposed model was compared with numerous advanced models, demonstrating its effectiveness. This section delves deeper into the discussion.

We employed three popular public datasets, DEAP, DREAMER and SEED, and used only four-channel EEG signals from the frontal polar region as inputs. This choice was based on two considerations. First, previous research

indicates that emotion-related EEG signals are predominantly found in the prefrontal lobe and lateral temporal lobe of the brain [17], [22], [40]. Second, the frontal polar region has less hair, reducing the likelihood of EEG signal interference from hair. Achieving high recognition accuracy with fewer EEG channels is a valuable exploration. Additionally, manual feature extraction requires specific domain knowledge and can disrupt the temporal characteristics of the original EEG signals, adding to the workload and potentially diminishing the dataset's usability.

With the rise of deep learning, self-attention mechanisms have garnered attention across various fields. However, our experiments revealed that Transformer-based models did not perform as expected with limited channels. Properly considering spatiotemporal characteristics can not only enhance recognition performance but also improve the interpretability of the EEG's temporal dependencies and spatial topology. Our two plug-and-play modules, MSTB and TSFB, are suited for different scenarios. From the experimental results, TSFB offered more significant benefits than MSTB. In simple data distribution scenarios, MSTB's improvement was minimal, whereas in complex environments, MSTB proved to be a valuable addition. Combining both modules enhances the model's generalization and robustness.

While MS-iMamba achieved impressive results using fewer channels, there are still several limitations. For instance, under the same experimental configuration, MS-iMamba's performance in cross-subject and cross-session scenarios was suboptimal. Given the challenges in acquiring large-scale EEG data, predicting unknown subjects' emotional categories using data from a few subjects remains challenging. However, our work suggests a potential method to preserve the data scale in EEG emotion recognition. In the future, we will continue exploring effective use of limited or incomplete data to improve MS-iMamba's performance in complex scenarios.

VI. CONCLUSION

This study introduces MS-iMamba, a novel model designed for EEG-based emotion recognition, integrating Multi-Scale Temporal Blocks (MSTB) and Temporal-Spatial Feature Blocks (TSFB). Our approach effectively captures multi-scale temporal features and spatiotemporal interactions, offering a robust alternative to traditional manual feature extraction methods. Comprehensive experiments conducted on three widely-used public datasets, DEAP, DREAMER, and SEED, demonstrate that MS-iMamba outperforms state-of-the-art models and achieves higher classification accuracy with fewer EEG channels.

Our results highlight the model's robustness and generalization capabilities. Notably, the combination of MSTB and TSFB enhances the model's performance, providing significant improvements over individual components. Despite the challenges in cross-subject and cross-session contexts, MS-iMamba's ability to achieve high accuracy with limited data channels underscores its potential for practical applications in real-world settings.

While MS-iMamba shows promise, it also faces limitations, particularly in handling the variability inherent in cross-subject and cross-session data. Future research will focus on further optimizing the model to handle these complexities and exploring the use of limited or incomplete data to enhance performance in more challenging scenarios.

In conclusion, MS-iMamba represents a significant advancement in EEG-based emotion recognition, offering a scalable, high-accuracy solution that balances the need for fewer data channels with robust performance. This work lays a foundation for future exploration in efficient and effective emotion recognition using EEG, with potential applications across various domains requiring precise emotional state detection.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (62176165), the Stable Support Projects for Shenzhen Higher Education Institutions (20220718110918001), the Natural Science Foundation of Top Talent of SZTU (GDRC202131), the Basic and Applied Basic Research Project of Guangdong Province (2022B1515130009), and the Special subject on Agriculture and Social Development, Key Research and Development Plan in Guangzhou (2023B03J0172).

REFERENCES

- [1] M. Jafari, A. Shoeibi, M. Khodatars, S. Bagherzadeh, A. Shalbaf, D. L. Garcia, J. M. Gorri, and U. R. Acharya, "Emotion recognition in eeg signals using deep learning methods: A review," *Computers in Biology and Medicine*, p. 107450, 2023.
- [2] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for eeg-based vigilance estimation," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 6627–6630.
- [3] 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.
- [4] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.
- [5] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "Sparsedgcn: Recognizing emotion from multichannel eeg signals," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 537–548, 2021.
- [6] P. Li, H. Liu, Y. Si, C. Li, F. Li, X. Zhu, X. Huang, Y. Zeng, D. Yao, Y. Zhang *et al.*, "Eeg based emotion recognition by combining functional connectivity network and local activations," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2869–2881, 2019.
- [7] H. Cui, A. Liu, X. Zhang, X. Chen, J. Liu, and X. Chen, "Eeg-based subject-independent emotion recognition using gated recurrent unit and minimum class confusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2740–2750, 2022.
- [8] L. Feng, C. Cheng, M. Zhao, H. Deng, and Y. Zhang, "Eeg-based emotion recognition using spatial-temporal graph convolutional lstm with attention mechanism," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5406–5417, 2022.
- [9] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition," *Information Processing & Management*, vol. 57, no. 3, p. 102185, 2020.
- [10] 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*, vol. 13, no. 3, pp. 1528–1540, 2020.
- [11] Z. Yang and H. Cao, "Decompose time and frequency dependencies: Multivariate time series physiological signal emotion recognition."
- [12] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.
- [13] W. K. Ngai, H. Xie, D. Zou, and K.-L. Chou, "Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources," *Information Fusion*, vol. 77, pp. 107–117, 2022.
- [14] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "Eeg-based emotion recognition via channel-wise attention and self attention," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 382–393, 2020.
- [15] C. Li, X. Lin, Y. Liu, R. Song, J. Cheng, and X. Chen, "Eeg-based emotion recognition via efficient convolutional neural network and contrastive learning," *IEEE Sensors Journal*, vol. 22, no. 20, pp. 19608–19619, 2022.
- [16] S. Liu, X. Wang, L. Zhao, B. Li, W. Hu, J. Yu, and Y.-D. Zhang, "3dcnn: A spatio-temporal convolution attention neural network for eeg emotion recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5321–5331, 2021.
- [17] X. Lin, J. Chen, W. Ma, W. Tang, and Y. Wang, "Eeg emotion recognition using improved graph neural network with channel selection," *Computer Methods and Programs in Biomedicine*, vol. 231, p. 107380, 2023.
- [18] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, and X. Chen, "Eeg-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network," *Knowledge-Based Systems*, vol. 205, p. 106243, 2020.
- [19] X. Deng, J. Zhu, and S. Yang, "Sfe-net: Eeg-based emotion recognition with symmetrical spatial feature extraction," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 2391–2400.
- [20] P. Gong, Z. Jia, P. Wang, Y. Zhou, and D. Zhang, "Astdf-net: Attention-based spatial-temporal dual-stream fusion network for eeg-based emotion recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 883–892.
- [21] C. Cheng, Z. Yu, Y. Zhang, and L. Feng, "Hybrid network using dynamic graph convolution and temporal self-attention for eeg-based emotion recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [22] L. Gong, M. Li, T. Zhang, and W. Chen, "Eeg emotion recognition using attention-based convolutional transformer neural network," *Biomedical Signal Processing and Control*, vol. 84, p. 104835, 2023.
- [23] L. Shen, M. Sun, Q. Li, B. Li, Z. Pan, and J. Lei, "Multiscale temporal self-attention and dynamical graph convolution hybrid network for eeg-based stereogram recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1191–1202, 2022.
- [24] 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.
- [25] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.
- [26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [27] M. Schuster and K. Nakajima, "Japanese and korean voice search," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 5149–5152.
- [28] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, J. Y. Zhang, and J. Zhou, "Timemixer: Decomposable multiscale mixing for time series forecasting," *arXiv preprint arXiv:2405.14616*, 2024.
- [29] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022.
- [30] P. Chen, Y. Zhang, Y. Cheng, Y. Shu, Y. Wang, Q. Wen, B. Yang, and C. Guo, "Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting," *arXiv preprint arXiv:2402.05956*, 2024.

- [31] H. Wang, L. Xu, A. Bezerianos, C. Chen, and Z. Zhang, "Linking attention-based multiscale cnn with dynamical gcn for driving fatigue detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2020.
- [32] Y. Jiang, S. Xie, X. Xie, Y. Cui, and H. Tang, "Emotion recognition via multiscale feature fusion network and attention mechanism," *IEEE Sensors Journal*, vol. 23, no. 10, pp. 10 790–10 800, 2023.
- [33] D. Li, J. Xu, J. Wang, X. Fang, and Y. Ji, "A multi-scale fusion convolutional neural network based on attention mechanism for the visualization analysis of eeg signals decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2615–2626, 2020.
- [34] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17 804–17 815, 2020.
- [35] N. Sesti, J. J. Garau-Luis, E. Crawley, and B. Cameron, "Integrating lstms and gnns for covid-19 forecasting," *arXiv preprint arXiv:2108.10052*, 2021.
- [36] J. Grigsby, Z. Wang, N. Nguyen, and Y. Qi, "Long-range transformers for dynamic spatiotemporal forecasting," *arXiv preprint arXiv:2109.12218*, 2021.
- [37] X. Chen and L. Sun, "Low-rank autoregressive tensor completion for multivariate time series forecasting," *arXiv preprint arXiv:2006.10436*, 2020.
- [38] A. Sharma and D. Kumar, "Classification with 2-d convolutional neural networks for breast cancer diagnosis," *Scientific Reports*, vol. 12, no. 1, p. 21857, 2022.
- [39] M. Jin and J. Li, "Graph to grid: Learning deep representations for multimodal emotion recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5985–5993.
- [40] C. Li, N. Bian, Z. Zhao, H. Wang, and B. W. Schuller, "Multi-view domain-adaptive representation learning for eeg-based emotion recognition," *Information Fusion*, vol. 104, p. 102156, 2024.
- [41] C. Cheng, W. Liu, L. Feng, and Z. Jia, "Dense graph convolutional with joint cross-attention network for multimodal emotion recognition," *IEEE Transactions on Computational Social Systems*, 2024.
- [42] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 11 121–11 128.
- [43] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.
- [44] 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.
- [45] S. Katsigianis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.
- [46] 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.
- [47] Y. Liu, H. Wu, J. Wang, and M. Long, "Non-stationary transformers: Exploring the stationarity in time series forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9881–9893, 2022.
- [48] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.
- [49] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
- [50] C. Li, Y. Hou, R. Song, J. Cheng, Y. Liu, and X. Chen, "Multi-channel eeg-based emotion recognition in the presence of noisy labels," *Science China Information Sciences*, vol. 65, no. 4, p. 140405, 2022.
- [51] T. Song, S. Liu, W. Zheng, Y. Zong, and Z. Cui, "Instance-adaptive graph for eeg emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2701–2708.
- [52] T. Song, S. Liu, W. Zheng, Y. Zong, Z. Cui, Y. Li, and X. Zhou, "Variational instance-adaptive graph for eeg emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 343–356, 2021.
- [53] F. Xu, D. Pan, H. Zheng, Y. Ouyang, Z. Jia, and H. Zeng, "Eescn: A novel spiking neural network method for eeg-based emotion recognition," *Computer methods and programs in biomedicine*, vol. 243, p. 107927, 2024.
- [54] C. Cheng, W. Liu, Z. Fan, L. Feng, and Z. Jia, "A novel transformer autoencoder for multi-modal emotion recognition with incomplete data," *Neural Networks*, vol. 172, p. 106111, 2024.

Multi-modal Mood Reader: Pre-trained Model Empowers Cross-Subject Emotion Recognition

Yihang Dong^{1,2[0000–1111–2222–3333]}, Xuhang Chen², Yanyan Shen^{1,2}, Michael Kwok-Po Ng³, Tao Qian⁴, and Shuqiang Wang^{1,2[0000–0003–1119–320X](✉)}

¹ University of Chinese Academy of Sciences

² Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

³ Department of Mathematics, Hong Kong Baptist University

⁴ Faculty of Innovation Engineering, Macau University of Science and Technology

sq.wang@siat.ac.cn

Abstract. Emotion recognition based on Electroencephalography (EEG) has gained significant attention and diversified development in fields such as neural signal processing and affective computing. However, the unique brain anatomy of individuals leads to non-negligible natural differences in EEG signals across subjects, posing challenges for cross-subject emotion recognition. While recent studies have attempted to address these issues, they still face limitations in practical effectiveness and model framework unity. Current methods often struggle to capture the complex spatial-temporal dynamics of EEG signals and fail to effectively integrate multimodal information, resulting in suboptimal performance and limited generalizability across subjects. To overcome these limitations, we develop a Pre-trained model based Multimodal Mood Reader for cross-subject emotion recognition that utilizes masked brain signal modeling and interlinked spatial-temporal attention mechanism. The model learns universal latent representations of EEG signals through pre-training on large scale dataset, and employs Interlinked spatial-temporal attention mechanism to process Differential Entropy(DE) features extracted from EEG data. Subsequently, a multi-level fusion layer is proposed to integrate the discriminative features, maximizing the advantages of features across different dimensions and modalities. Extensive experiments on public datasets demonstrate Mood Reader’s superior performance in cross-subject emotion recognition tasks, outperforming state-of-the-art methods. Additionally, the model is dissected from attention perspective, providing qualitative analysis of emotion-related brain areas, offering valuable insights for affective research in neural signal processing.

Keywords: EEG-based emotion recognition · Pre-trained Model · spatial-temporal attention · masked brain signal modeling

1 Introduction

Brain-computer interface (BCI) systems have long been an aspirational goal for researchers in the fields of computer science, neuroscience, and psychology.

The envisioned maturation of BCI technology is expected to significantly expand human sensory, cognitive, and operational capabilities, offering unprecedented depth and breadth in human-machine interaction. However, truly efficient human-machine interaction relies not solely on the machine's ability to interpret and execute human commands, but more critically, on the sensitive detection and accurate recognition of users' implicit emotional states. Consequently, the task of emotion recognition has naturally emerged as a key research area.

Although emotion recognition methods based on various physiological signals each have their unique characteristics, they predominantly face challenges related to the complexity of signal collection, difficulties in data processing, and high costs. Therefore, non-invasive EEG, with its relatively low cost, convenient signal collection, superior signal representation capability, and non-harmful nature to subjects, has rapidly become a primary research focus in the field of emotion recognition. The array of electrodes placed on the scalp effectively collects signals reflecting brain electrical activity. Through precise analysis and processing of these signals, an individual's emotional state can be effectively revealed.

Non-invasive EEG signals are not without flaws. The unique natural physiological and anatomical structures of each individual introduce various degrees and aspects of noise interference into the measured EEG, imparting non-stationary characteristics to it. Additionally, issues such as the non-Euclidean distribution of multi-channel EEG electrodes based on biological topography collectively impact the accuracy of cross-subject emotion recognition tasks. Researchers have attempted to tackle these challenges from different directions. Transfer learning, as an indirect approach, has been utilized to migrate emotion recognition models, originally trained and adapted for existing subjects, to new individuals, aiming to minimize the EEG differences between the source and target domains [1,2]. Although this approach has indeed achieved certain recognition effects, it does not fundamentally solve the problem of cross-subject emotion recognition. Considering the graph-like topological structure of EEG channels and the rapid development of Graph Neural Networks (GNN), a surge of cross-subject emotion recognition methods based on GNN has emerged [3,4], attempting to capture the local and global relationships among EEG channels. Similarly, other methods have also achieved certain yet limited improvements in recognition accuracy [5,6].

Motivated by the recent emergence of pre-trained models and their outstanding performance in downstream tasks [7,8,9,10,11], we recognized that the high-dimensional semantic information of EEG extracted by encoders trained on large-scale subject-independent datasets may contain global generic representations beneficial for emotion recognition tasks. Simultaneously, considering the DE feature, which has been proven to be the most effective individual computational characteristic for EEG-based emotion recognition tasks [12,13,14], as well as eye movement features often used as an additional modality to aid emotion recognition tasks and confirmed to improve recognition accuracy [15], we propose Mood Reader, a novel multi-modal and cross-scale fusion model for cross-subject emotion recognition that integrates global generic representations of EEG and

the spatio-temporal interaction information in specific DE features. Specifically, our contributions are as follows:

1. We propose an emotion recognition model architecture that integrates multi-modal and cross-scale information, demonstrating exceptional performance in cross-subject recognition tasks. This architecture also proves that encoders pre-trained on large-scale EEG data possess the ability to learn emotion-related features to a certain extent.
2. We have designed an attention-based interlinked spatio-temporal module for learning the compensatory relationships between spatio-temporal information, which aids in the fusion of spatio-temporal features.
3. Supported by extensive experiments, we provide a biologically plausible interpretation of emotion recognition research based on EEG and make reasonable hypotheses.

2 Related Works

2.1 EEG-Based Emotion Recognition

Research on EEG-based emotion recognition has attracted significant attention from researchers in recent years. EEG signals, due to their inherent attribute of directly measuring the electrical activity on the scalp surface to capture changes in brain neural activity, have almost become the most powerful data type for emotion recognition. With the development of deep learning technologies that automatically extract features from data, there has been increasing focus on their application in critical areas such as computer vision, natural language processing, and emotion recognition. Given that EEG signals are essentially multi-channel time series signals, to reasonably utilize this nature, Li et al. proposed a BiLSTM network framework based on multimodal attention, which is used to learn the best temporal characteristics, and input the learned deep features into the DNN to predict the emotional output probability of each channel [16]. Since emotions are the comprehensive result of the human body's response to external stimuli [17], researchers naturally began to use attention mechanisms, similarity coordination constraints, and other multimodal fusion methods to integrate single-modal features extracted from EEG, eye movement signals, facial expressions, etc., from different perspectives for emotion recognition research [18,19,20,21,22,23]. The ablation studies of these researches powerfully validate the correctness of the direction of multimodal data fusion in the field of emotion recognition.

2.2 Masked Brain Signal Modeling

With the rapid development of self-supervised pre-training models in the fields of computer vision and natural language processing, researchers have migrated this technology, which can learn generic knowledge representations for target tasks, to multiple fields including brain signal decoding. MBSM, proposed by

Chen et al. [7], is a self-supervised learning model for large-scale fMRI datasets, which helps its encoder learn the general representation in fMRI signals through the learning process of repeatedly reconstructing complete data from unmasked fMRI signals, further adapting the encoder to different downstream tasks using simple fine-tuning techniques [7,24]. Meanwhile, Bai et al. successfully overcame the inherent variability and noise of EEG data by deeply mining the semantics of EEG signals over time, migrated this technology to EEG, and applied it to downstream tasks such as decoding high-resolution images from brain activity [25].

2.3 Spatial-Temporal Attention Mechanism

Complex neural activities are often effectively achieved by the synergistic collaborative processing of multiple sets of continuous neural signals across various distinct neural regions. [26,27,28,29,30,31,32,33]. Therefore, for research fields including brain decoding and neural information processing, it is crucial to simultaneously capture the temporal and spatial information in brain signal data, for instance, fMRI and EEG. Since the attention mechanism was proposed, various application fields of deep learning, including but not limited to the field of emotion recognition based on EEG signals, have made great progress, and many excellent spatiotemporal information extraction modules have been deduced [34,35,36]. In the realm of EEG-based emotion recognition, the Spatial-Temporal Attention (STA) mechanism emerges as a notable innovation for enhancing the interpretability and performance of deep learning models. This type of architecture ingeniously integrates spatial and temporal dimensions of EEG signals through parallel attention pathways, enabling the model to concurrently learn spatial correlations across EEG channels and temporal dependencies within signal sequences. Li et al. employed the spatio-temporal combination network R2G-STNN, which contains local-global feature combining, to extract the intrinsic information of EEG signals [37]. In order to mine the spatiotemporal information related to emotional judgment, Gong et al. designed a stacked parallel spatial and temporal attention streams to respectively extract the spatial features and temporal features of the specially processed EEG signals [38]. Although previous researchers have obtained satisfactory results, most of them have ignored the interaction and mutual compensation between spatial information and temporal information in EEG signals. In these network streamlines, the two kinds of information are often ignored. Parallel offload processing, which is contrary to the processing flow of neural signals in complex neural activities, may not be enough to handle more complex neural signal processing tasks.

3 Methodology

3.1 Overview

Mood Reader is an emotion recognition model that accommodates multi-modal, cross-scale information and successfully integrates these features, its overall architecture is illustrated in Figure 1. The model encompasses three distinct types

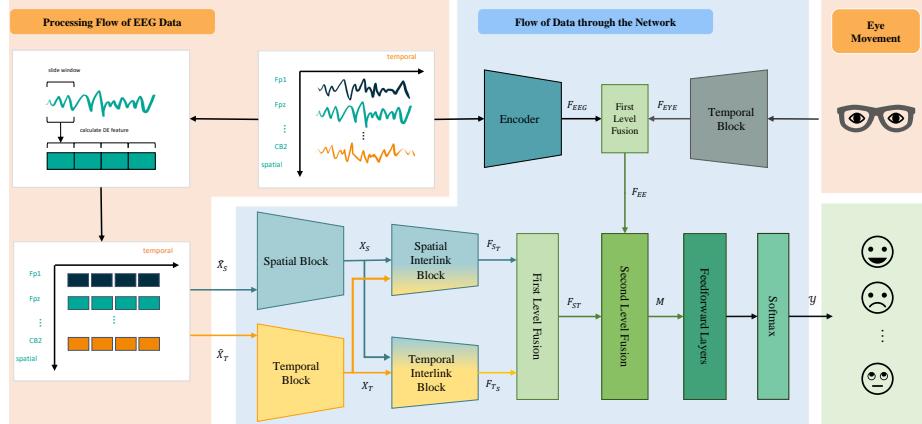


Fig. 1: The overall architecture of our proposed model and the way related data flows in it.

of input. EEG monitoring data, subjected to simple preprocessing, are encoded by an encoder pre-trained on a large-scale dataset, resulting in outputs that contain rich semantic representations. An attention-based interlinked spatial-temporal mechanism captures the intrinsic spatio-temporal information from DE features extracted from EEG data. Additionally, a set of similar temporal attention blocks analyzes corresponding eye movement features, aiming to complement the shortcomings of EEG data. The acquired features are progressively fused in a sequential manner, ensuring that the model genuinely learns the comprehensive complementarity between different modal and scale information, and utilizes it for emotion recognition tasks.

3.2 MBSM Based EEG General Representation Learning

Due to the inherent brain differences among subjects and external noise affecting the signals collected by non-invasive EEG, we adopted a pre-training technique known as masked brain signal modeling, which has been proven effective multiple times, to learn meaningful and contextually rich general knowledge representations from cross-subject, noisy, large-scale EEG data [25,7,24]. Specifically, we completed this task by training an autoencoder-decoder with an asymmetric architecture similar to that in [25] on the EEG Motor Movement/Imagery Dataset [39]. In this model, the temporal signals of EEG data were divided into tokens of a specific size, where a larger ratio of tokens would be randomly masked, and the architecture-simple decoder had to reconstruct the EEG data using the remaining unmasked tokens arrangement, combined with semantically rich embeddings outputted by the encoder after processing the original EEG data. The performance of masked reconstruction improves and reaches a peak when the mask ratio hits 75% [25,7,24]. Consequently, by removing the decoder from this trained model, we obtained an encoder with excellent capability in extracting general EEG representations.

3.3 Attention Based Interlinked Spatial-Temporal Mechanism

Deep learning analyses of neural processes typically begin with various types of feature extraction. For complex neural activities like EEG-based emotion recognition, which exhibit high spatiotemporal continuity, it's crucial to accurately unearth intrinsic spatial and temporal features and their interrelations to obtain more valuable information. To address this challenge, this section introduces the interlinked spatial-temporal attention module for processing DE features, comprised of multiple parallel spatiotemporal blocks and interactive spatiotemporal blocks (as depicted in Figure.), these blocks collectively facilitate the extraction of spatio-temporal information and enable the communication and complementation between the extracted spatial and temporal features.

Spatial and Temporal Representation of DE features In the domain of EEG-based emotion recognition, the DE feature, which quantifies the variability of EEG signals, has been proven to be the most effective feature, capturing brain activities related to emotions. It itself has channel-related explicit spatial information and implicit temporal information compressed within a single sliding window. In order to make the spatiotemporal information in the DE feature more balanced, we expand the number of sliding windows in the DE feature to the explicit temporal dimension, obtaining $\hat{X} \in \mathbb{R}^{N \times F \times C}$, where N denotes the number of sliding windows involved in the DE feature computation. For further dimension transformation, we get $\hat{X}_S \in \mathbb{R}^{C \times (N \cdot F)}$ for spatial information representation and $\hat{X}_T \in \mathbb{R}^{N \times (C \cdot F)}$ for temporal information representation.

Parallel Spatiotemporal Feature Extraction To capture the dynamically varying key information, we apply dedicated spatial and temporal attention blocks to the differentiated DE feature spatial representation \hat{X}_S and temporal representation \hat{X}_T , respectively. Specifically, for the spatial representation $\hat{X}_S \in \mathbb{R}^{C \times (N \cdot F)}$, an initial layer normalization is employed to yield X'_S , which effectively mitigates the internal covariate shift within the spatial representation data, thereby maintaining the stability of its distribution, as follows,

$$X'_S = \text{LayerNorm}(\hat{X}_S) \quad (1)$$

After the normalization, multi-head attention (MHA) computation is implemented on X'_S . Within head i , X'_S is processed by three separate linear networks, transforming the input into different representational spaces to obtain the corresponding query Q_S^i , key K_S^i , and value V_S^i , denoted as,

$$Q_S^i = X'_S W_S^{q_i}, \quad K_S^i = X'_S W_S^{k_i}, \quad V_S^i = X'_S W_S^{v_i} \quad (2)$$

Where $W_S^{q_i}$, $W_S^{k_i}$, $W_S^{v_i}$ are the learnable network parameters respectively.

Based on the attention calculation method, the attention output for each head i is obtained as $A_S^i = \text{Attention}(Q_S^i, K_S^i, V_S^i)$. With some processing, the DE feature space representation's MHA output A_S can be obtained as $A_S =$

$\text{Concat}(A_S^1, A_S^2, \dots, A_S^h)W_S^{\text{MHA}}$, where h is the number of attention heads, and W_S^{MHA} is a linear mapping weight. The attention calculation is described as,

$$\text{Attention}(Q_S^i, K_S^i, V_S^i) = \text{softmax} \left(\frac{Q_S^i K_S^{i^T}}{\sqrt{d_k}} \right) V_S^i \quad (3)$$

where d_k represents the dimension of the key.

At the conclusion of the spatial attention block, a residual connection is introduced by adding the dropout-processed MHA output A_S to X'_S . This combined result is then subjected to another layer normalization to produce the final spatial representation of the DE feature X_S , which is within the space $\mathbb{R}^{C \times (N \cdot F)}$. The calculation process of this part is as follows,

$$X_S = \text{LayerNorm}(\text{Dropout}(A_S) + X'_S) \quad (4)$$

Similarly, we process $\hat{X}_T \in \mathbb{R}^{N \times (C \cdot F)}$ with a structurally analogous temporal attention block to obtain the temporal representation of the DE feature $X_T \in \mathbb{R}^{N \times (C \cdot F)}$

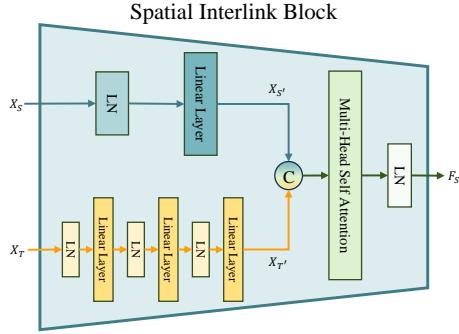


Fig. 2: Overview of the spatial interlink block, the temporal feature X_T undergoes multiple transformations to align with the spatial feature X_S , and after concatenation, the interlink process is completed through multi-head attention computation.

Interlink Between Spatio-Temporal Block For the spatial features X_S and temporal features X_T , which are processed by the spatio-temporal attention block, they are fed into the corresponding interlink block to facilitate the intersection of primary and secondary dimensions. Specifically, the interlinked spatial block will receive both X_S and X_T . However, for the primary information X_S , which has already achieved a high degree of spatial information integration after being processed by the spatial attention block, it only requires a single linear transformation to become X'_S , in order to preserve its existing information.

As for X_T , it undergoes multiple transformations in an attempt to align with the spatial feature dimensions, resulting in $X_{T'}$. Subsequently, we concatenate these two feature parts and perform MHA calculation to obtain the spatial features F_{S_T} that are interlinked with the temporal dimension, as shown in Figure 2. Similarly, we can also obtain the temporal features F_{T_S} that are interlinked with the spatial dimension.

3.4 Multi-Level Score Filtering for Feature Fusion

Previous sections presented outputs from various modules, each designed for emotion recognition through different pathways, yielding high-dimensional features. To address potential redundancy and the needs of multimodal cross-scale fusion, we introduce a multi-level fusion layer based on attention mechanisms. This module synergizes and refines features by highlighting relevant information and filtering out redundancy, thus improving multimodal emotion recognition efficacy.

Prior to the commencement of the fusion process, the MBSM latent representation, spatial-temporal representation, and eye movement representation, are projected onto a unified dimensional space through a series of transformations, starting with layer normalization, followed by flattening of the feature vectors, and culminating in a linear transformation. As a result of these operations, a unified feature representation $F \in \mathbb{R}^{D_{\text{unified}}}$ is obtained, where D_{unified} denotes the dimensionality of the unified feature space.

In the first level of fusion, specifically for the interlinked spatial-temporal features F_{S_T} and F_{T_S} , preprocessing is executed utilizing a linear layer and layer normalization, which preserves their dimensional attributes. Subsequently, inspired by [18,40,41], a simplified cross-attention mechanism is employed to delineate the intrinsic spatial-temporal relationships between the two, thereby augmenting the interactivity of the internal information. Specifically, the attention scores Score_s and Score_t , which represent the spatial features taking into account temporal features and time features taking into account spatial features respectively, are transformed using the softmax function to ascertain the corresponding fusion weights c_s and c_t ,

$$\text{Score}_s = (F_{S_T} W), \text{Score}_t = (F_{T_S} W) \quad (5)$$

$$c_s, c_t = \text{softmax}(\text{Score}_s, \text{Score}_t), \quad (6)$$

where $W_S, W_T \in \mathbb{R}^{D_{\text{unified}} \times D_{\text{unified}}}$ is a learnable weight matrix, the process of obtaining c_t is analogous to this. The spatial-temporal fusion features F_{ST} are then calculated by,

$$F_{ST} = c_s F_{S_T} + c_t F_{T_S} \quad (7)$$

The processing of the MBSM latent representation and eye movement representation, corresponding to F_{EEG} and F_{EYE} respectively, is conducted in a manner akin to the aforementioned methodology. This approach yields an additional fused feature F_{EE} , computed as a weighted sum: $F_{EE} = c_{eeg} F_{EEG} + c_{eye} F_{EYE}$,

where c_{eeg} and c_{eye} represent the fusion weights derived from the respective attention scores. In the terminal fusion layer, an integration is requisite for the synthesized features F_{ST} and F_{EE} , which respectively represent the interlinked spatio-temporal information emanating from the DE feature and the comprehensive information spanning multiple modalities and scales. The features that have undergone layer normalization are concatenated and subsequently processed through self-attention computation, yielding the final integrated feature representation M ,

$$M = \text{Attention}(\text{Concat}(F_{ST}, F_{EE})) \quad (8)$$

For the feature M , an initial batch normalization is employed, followed by the deployment of a classifier comprised of three linear layers and a softmax function, which outputs predictions for the emotional labels y . The discrepancy between the predicted emotional labels and the true emotional labels \hat{y} is quantified using the cross-entropy loss function,

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C \hat{y}_{ic} \log(y_{ic}) \quad (9)$$

wherein N represents the batch size, and C designates the count of label categories.

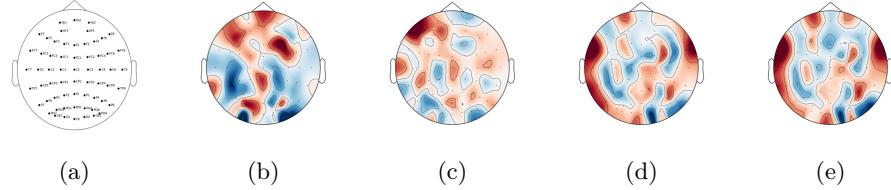


Fig. 3: Attention visualization. We visualized the model’s attention weights at different moments during the training process, which allows for an intuitive understanding of how the model’s preference for EEG signals monitored by electrodes at various locations evolves over time (b, c, d, e). (a) presents the layout of the utilized 62-channel electrode placement.

3.5 Interpretability

In order to provide a biologically plausible interpretation and inference for the research on EEG-based emotion recognition, as well as to substantiate the efficacy of the proposed model, we have conducted multiple visualizations of the attention outputs during the encoding process of EEG signals from test samples, projecting them back onto the electrode location map. This approach has allowed for a more intuitive validation of the brain regions associated with the task of emotion recognition, thereby enabling more insightful deductions.

4 Experiment Result and Analysis

4.1 Datasets and Pre-processing

Extensive experiments have been conducted on two public datasets, SEED and SEED-V. The SEED dataset includes EEG monitoring data and corresponding emotional labels (sad, happy, and neutral) from 15 subjects. Each subject completed 3 sessions, with each session comprising 15 trials, resulting in a total of 45 trials. Similarly, SEED-V was completed by 16 subjects, each participating in 45 trials, and includes EEG data along with corresponding labels for five emotions (disgust, fear, sad, happy, and neutral).

In the preprocessing of raw EEG data, a sequence involving a bandpass filter with cutoff frequencies of 0.1Hz and 70Hz, followed by a notch filter at 50Hz, was implemented. Subsequently, the sampling frequency was reduced to 200Hz from its original rate. For EEG segments corresponding to trials of varied lengths, a 4s non-overlapping Hanning window was utilized for segmentation in reverse order. The Short Time Fourier Transform (STFT) was then applied to calculate the DE feature across five frequency domains. Every four sliding window calculations were grouped together to extract EEG signals requisite for the pretrained model.

4.2 Baseline Model and Settings

We conducted subject-independent experiments on the SEED and SEED-V datasets using baseline models including DGCNN [42], RGNN [4], SOGNN [43] and BFE-Net [44]. Notably, baseline models that were trained using a single modality were explicitly annotated. The experimental framework utilized was PyTorch, with the GPU being NVIDIA A800 80GB PCIe. Furthermore, for each experiment, the data were randomly divided into training and testing sets in an 8:2 ratio. The model performance was evaluated based on the average accuracy and variance on the testing set.

Table 1: Subject-independent classification performance (Acc/Std%) on SEED and SEED-V, where SWC represents whether to combine DE features in sliding window order.

Method	Modality			SEED		SEED-V	
	DE Feature w/o SWC	DE Feature w SWC	Raw EEG Eye Movement	Acc	Std	Acc	Std
SVM	✓			56.73	16.29	34.45	13.67
DGCNN	✓			79.95	09.02	-	-
RGNN	✓			85.30	06.72	66.28	16.71
SOGNN	✓			86.81	05.79	74.53	07.90
BFE-Net	✓			92.29	04.65	-	-
SVM		✓		70.23	10.42	56.49	12.51
Mood reader	✓		✓	91.65	05.42	78.74	08.97
Mood reader	✓		✓	93.12	04.75	84.36	05.23

4.3 Results

Experiment Result Table 1 presents the experimental results of the baseline model and our method on the SEED and SEED-V datasets, with annotations regarding the categories of data utilized. The results demonstrate the consistently superior classification performance of Mood Reader across different datasets. Furthermore, they validate the efficacy of the sequentially combined DE features through a sliding window approach in the task of emotion recognition, as well as the correctness of the multimodal cross-scale information fusion strategy.

Interpretation The results of attention visualization are summarized in Figure 3. It can be observed that as the emotional recognition capability improves, the network’s attention on EEG signals gradually shifts from a scattered global distribution to concentrated attention on specific regions. These brain regions include the frontal lobe area, areas of the left and right temporal lobes, and a small portion of the parietal lobe, which has been proven to be closely related to the generation and processing of emotions [45].

Additionally, we noticed that there are also small areas within the occipital lobe, primarily responsible for visual information processing, that exhibit significant attention. Given that a substantial part of the stimuli in the experimental datasets SEED and SEED-V comes from visual stimuli in videos, we have reasonable grounds to propose the hypothesis that “visually encoded information in the human brain is directly involved in emotion generation” to a certain extent. This hypothesis also represents the holistic view that various parts of the brain participate in different functions and collectively process complex information [46].

4.4 Ablation Studies

To substantiate the effectiveness of the employed modules, we conducted ablation experiments on the SEED-V dataset using a stepwise stacking approach for the modules, with the specific experimental details as follows, and the results are depicted in Figure 4.

1. STB+CF: SWC DE with spatial-temporal block + concatenation fusion
2. STIB+CF: SWC DE with spatial-temporal interlinked block + concatenation fusion
3. STIB+Encoder+CF: SWC DE with spatial-temporal interlinked block + pre-trained encoder + concatenation fusion
4. STIB+Encoder+MLF: SWC DE with spatial-temporal interlinked block + pre-trained encoder + multi-level fusion
5. STIB+Eye+CF: SWC DE with spatial-temporal interlinked block + eye movement + concatenation fusion
6. STIB+Eye+MLF: SWC DE with spatial-temporal interlinked block + eye movement + multi-level fusion

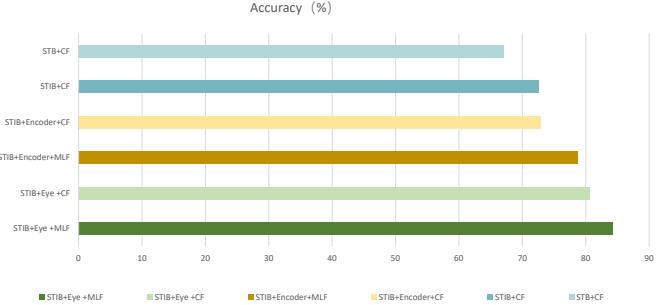


Fig. 4: The results of the ablation studies, conducted through progressive stacking of modules.

5 Conclusion

In this paper, we present the Mood Reader, a novel multimodal cross-scale fusion model for cross-subject emotion recognition based on EEG signals. Our model effectively integrates masked brain signal modeling for learning universal latent representations and an interlinked spatial-temporal attention mechanism to capture the complex dynamics of EEG signals. The multi-level fusion layer maximizes the advantages of features across different dimensions and modalities, leading to superior performance in cross-subject emotion recognition tasks. Furthermore, the model’s interpretability, achieved through attention visualization, provides valuable insights into emotion-related brain areas, contributing to the understanding of neural processes underlying emotions. In conclusion, Mood Reader represents a significant step forward in cross-subject EEG-based emotion recognition, leveraging multimodal cross-scale fusion and advanced attention mechanisms.

Acknowledgement

This work was supported in part by the National Natural Science Foundations of China under Grant 62172403, the Distinguished Young Scholars Fund of Guangdong under Grant 2021B1515020019. M. Ng’s research is supported in part by the HKRGC GRF 17201020 and 17300021, HKRGC CRF C7004-21GF, and Joint NSFC and RGC N-HKU769/21.

References

1. Li, J., Qiu, S., Shen, Y.Y., Liu, C.L., He, H.: Multisource transfer learning for cross-subject eeg emotion recognition. *IEEE transactions on cybernetics* **50**(7), 3281–3293 (2019)

2. Yan, H., Zhang, H., Shi, J., Ma, J., Xu, X.: Inspiration transfer for intelligent design: A generative adversarial network with fashion attributes disentanglement. *IEEE Transactions on Consumer Electronics* (2023)
3. Li, Y., Chen, J., Li, F., Fu, B., Wu, H., Ji, Y., Zhou, Y., Niu, Y., Shi, G., Zheng, W.: Gmss: Graph-based multi-task self-supervised learning for eeg emotion recognition. *IEEE Transactions on Affective Computing* (2022)
4. Zhong, P., Wang, D., Miao, C.: Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing* **13**(3), 1290–1301 (2020)
5. Li, J., Hua, H., Xu, Z., Shu, L., Xu, X., Kuang, F., Wu, S.: Cross-subject eeg emotion recognition combined with connectivity features and meta-transfer learning. *Computers in biology and medicine* **145**, 105519 (2022)
6. Wang, S., Shen, Y., Zeng, D., Hu, Y.: Bone age assessment using convolutional neural networks. In: 2018 International conference on artificial intelligence and big data (ICAIBD). pp. 175–178. IEEE (2018)
7. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22710–22720 (2023)
8. Ortega Caro, J., Oliveira Fonseca, A.H., Averill, C., Rizvi, S.A., Rosati, M., Cross, J.L., Mittal, P., Zappala, E., Levine, D., Dhadapkar, R.M., et al.: Brainlm: A foundation model for brain activity recordings. *bioRxiv* pp. 2023–09 (2023)
9. Luo, S., Chen, X., Chen, W., Li, Z., Wang, S., Pun, C.M.: Devignet: High-resolution vignetting removal via a dual aggregated fusion transformer with adaptive channel expansion. In: AAAI Conference on Artificial Intelligence. pp. 4000–4008 (March 2024)
10. Li, Z., Chen, X., Pun, C.M., Cun, X.: High-resolution document shadow removal via a large-scale real-world dataset and a frequency-aware shadow erasing net. In: International Conference on Computer Vision (ICCV). pp. 12449–12458 (October 2023)
11. Li, Z., Chen, X., Wang, S., Pun, C.M.: A large-scale film style dataset for learning multi-frequency driven film enhancement. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 1160–1168 (August 2023)
12. Du, X., Ma, C., Zhang, G., Li, J., Lai, Y.K., Zhao, G., Deng, X., Liu, Y.J., Wang, H.: An efficient lstm network for emotion recognition from multichannel eeg signals. *IEEE Transactions on Affective Computing* **13**(3), 1528–1540 (2020)
13. Tao, W., Li, C., Song, R., Cheng, J., Liu, Y., Wan, F., Chen, X.: Eeg-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing* **14**(1), 382–393 (2020)
14. Shen, X., Liu, X., Hu, X., Zhang, D., Song, S.: Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing* (2022)
15. Liu, W., Qiu, J.L., Zheng, W.L., Lu, B.L.: Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems* **14**(2), 715–729 (2021)
16. Li, C., Bao, Z., Li, L., Zhao, Z.: Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition. *Information Processing & Management* **57**(3), 102185 (2020)
17. LeDoux, J.E.: Cognitive-emotional interactions in the brain. *Cognition & Emotion* **3**(4), 267–289 (1989)

18. Jiang, W.B., Liu, X.H., Zheng, W.L., Lu, B.L.: Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5975–5984 (2023)
19. Vazquez-Rodriguez, J., Lefebvre, G., Cumin, J., Crowley, J.L.: Emotion recognition with pre-trained transformers using multimodal signals. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–8. IEEE (2022)
20. Jia, Z., Lin, Y., Wang, J., Feng, Z., Xie, X., Chen, C.: Hetemotionnet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1047–1056 (2021)
21. Ma, J., Tang, H., Zheng, W.L., Lu, B.L.: Emotion recognition using multimodal residual lstm network. In: Proceedings of the 27th ACM international conference on multimedia. pp. 176–183 (2019)
22. Chaparro, V., Gomez, A., Salgado, A., Quintero, O.L., Lopez, N., Villa, L.F.: Emotion recognition from eeg and facial expressions: a multimodal approach. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 530–533. IEEE (2018)
23. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: Emotionmeter: A multimodal framework for recognizing human emotions. IEEE transactions on cybernetics **49**(3), 1110–1122 (2018)
24. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. Advances in Neural Information Processing Systems **36** (2024)
25. Bai, Y., Wang, X., Cao, Y.p., Ge, Y., Yuan, C., Shan, Y.: Dreamdiffusion: Generating high-quality images from brain eeg signals. arXiv preprint arXiv:2306.16934 (2023)
26. Yang, E., Milisav, F., Kopal, J., Holmes, A.J., Mitsis, G.D., Misic, B., Finn, E.S., Bzdok, D.: The default network dominates neural responses to evolving movie stories. Nature communications **14**(1), 4197 (2023)
27. Rollo, J., Crawford, J., Hardy, J.: A dynamical systems approach for multiscale synthesis of alzheimer’s pathogenesis. Neuron **111**(14), 2126–2139 (2023)
28. You, S., Lei, B., Wang, S., Chui, C.K., Cheung, A.C., Liu, Y., Gan, M., Wu, G., Shen, Y.: Fine perceptive gans for brain mr image super-resolution in wavelet domain. IEEE transactions on neural networks and learning systems (2022)
29. Gong, C., Jing, C., Chen, X., Pun, C.M., Huang, G., Saha, A., Nieuwoudt, M., Li, H.X., Hu, Y., Wang, S.: Generative ai for brain image computing and brain network computing: a review. Frontiers in Neuroscience **17**, 1203104 (2023)
30. Wang, S., Wang, H., Cheung, A.C., Shen, Y., Gan, M.: Ensemble of 3d densely connected convolutional network for diagnosis of mild cognitive impairment and alzheimer’s disease. Deep learning applications pp. 53–73 (2020)
31. Hu, B., Zhan, C., Tang, B., Wang, B., Lei, B., Wang, S.Q.: 3-d brain reconstruction by hierarchical shape-perception network from a single incomplete image. IEEE Transactions on Neural Networks and Learning Systems (2023)
32. Pan, J., Lei, B., Shen, Y., Liu, Y., Feng, Z., Wang, S.: Characterization multimodal connectivity of brain network by hypergraph gan for alzheimer’s disease analysis. In: Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4. pp. 467–478. Springer (2021)

33. Wang, S.Q.: A variational approach to nonlinear two-point boundary value problems. *Computers & Mathematics with Applications* **58**(11-12), 2452–2455 (2009)
34. Cherian, A., Wang, J., Hori, C., Marks, T.: Spatio-temporal ranked-attention networks for video captioning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1617–1626 (2020)
35. Ahn, D., Kim, S., Hong, H., Ko, B.C.: Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3330–3339 (2023)
36. Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., Tao, D.: Transvod: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
37. Li, Y., Zheng, W., Wang, L., Zong, Y., Cui, Z.: From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing* **13**(2), 568–578 (2019)
38. Gong, P., Jia, Z., Wang, P., Zhou, Y., Zhang, D.: Astdf-net: Attention-based spatial-temporal dual-stream fusion network for eeg-based emotion recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 883–892 (2023)
39. Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R.: Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering* **51**(6), 1034–1043 (2004)
40. Zuo, Q., Wu, H., Chen, C.P., Lei, B., Wang, S.: Prior-guided adversarial learning with hypergraph for predicting abnormal connections in alzheimer’s disease. *IEEE Transactions on Cybernetics* (2024)
41. Zuo, Q., Lei, B., Shen, Y., Liu, Y., Feng, Z., Wang, S.: Multimodal representations learning and adversarial hypergraph fusion for early alzheimer’s disease prediction. In: Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4. pp. 479–490. Springer (2021)
42. Song, T., Zheng, W., Song, P., Cui, Z.: Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing* **11**(3), 532–541 (2018)
43. Li, J., Li, S., Pan, J., Wang, F.: Cross-subject eeg emotion recognition with self-organized graph neural network. *Frontiers in Neuroscience* **15**, 611653 (2021)
44. Zhang, J., Hao, Y., Wen, X., Zhang, C., Deng, H., Zhao, J., Cao, R.: Subject-independent emotion recognition based on eeg frequency band features and self-adaptive graph construction. *Brain Sciences* **14**(3), 271 (2024)
45. Dolcos, F., LaBar, K.S., Cabeza, R.: Interaction between the amygdala and the medial temporal lobe memory system predicts better memory for emotional events. *Neuron* **42**(5), 855–863 (2004)
46. Sporns, O.: Structure and function of complex brain networks. *Dialogues in clinical neuroscience* **15**(3), 247–262 (2013)