ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096511

ELASTIC GRAPH TRANSFORMER NETWORKS FOR EEG-BASED EMOTION RECOGNITION

Wei-Bang Jiang[1], Xu Yan[2], Wei-Long Zheng[1], Bao-Liang Lu[1]†

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2]Department of Linguistics, University of Washington, U.S.A.

ABSTRACT

Electroencephalogram (EEG) has been applied in emotion recognition due to excellent temporal resolution with less competitive spatial resolution. This leads to the consequence that the majority of EEG-based emotion recognition models emphasize on exploiting temporal features while ignoring the efficient information provided by spatial resolution. To extract more informative representations, we propose an elastic Graph Transformer network for emotion recognition (EmoGT) inspired by the advantages of Transformer in time-series analysis and the superior performance of graph convolutional networks in topological analysis. Moreover, it is able to be flexibly expanded to cope with multimodal inputs by employing specially designed structures. Experimental results on 3 public datasets demonstrate that our models outperform the state-of-the-art results by 3% on average in both single and multimodal cases, indicating the effectiveness of utilizing temporal and spatial information simultaneously.

***Index Terms*—** EEG, eye movements, emotion recognition, graph transformer

1. INTRODUCTION

Emotion recognition (ER) is the salient milestone in realization of emotion intelligence of AI. It serves as the foundation for a wide range of potential applications in everyday life, such as medical diagnosis, intelligent education, entertainment, etc. [1]. Many physiological signals are adopted as inputs to measure emotions, including speech, facial expression, eye movements, and EEG signals. Among all these modalities, the noninvasive EEG-based ER has gained increasing attention due to the efficacy of the brain in emotion processing from the evidence of cognitive neuroscience [2].

Previous studies evaluate brain imaging techniques on two dimensions, i.e., the spatial and temporal resolutions [3]. Under these criteria, EEG is considered to have a high temporal resolution in milliseconds but a relatively low spatial resolution in centimeters. Because of the excellent performance in temporal resolution, researchers usually take advantage of its merit and therefore extract EEG features from the temporal domain [4] and build temporal models.

There is considerable literature on EEG-based ER tasks using temporal models. The deep learning methods, ranging from convolutional neural networks (CNN) [5], recurrent neural networks (RNN) [6], to long short-term memory networks (LSTM) [7, 8], have made great strides in their performances on EEG-based ER. Though these models have yielded good results, they lack parallelization with low efficacy. To address this issue, therefore, we leverage Transformer [9] structure to capture the temporal information of EEG signals, which reduces complexity and computational cost by avoiding recursion and ingesting long sequential inputs once as a whole.

Since most studies are devoted to research in the temporal domain, the spatial information conveyed by EEG signals has not been fully utilized. Several studies applied CNN to extract the spatial connections. Nevertheless, CNN can only capture local information based on Euclidean distance, resulting in the unsatisfactory performance. It is the essence of emotion processing in brain that inspires us to think from its neurological support in model design. It has been proved that the connections between different regions in the circuits level, rather than local activity patterns, vary between different emotions and contribute to emotion detection [10]. The emerging graph convolutional networks (GCN) are able to flexibly process graph-structured data like EEG by taking each electrode as a node of the graph. Some variants of GCN have been applied on EEG to explore the inter-channel information, taking the regularized graph neural network proposed by Zhong *et al.* [11], Song *et al.* [12], and Li *et al.* [13] as examples. However, how to simultaneously extract spatial and temporal information in GCN has not been fully explored.

To fully exploit both spatial and temporal information of EEG in ER, we propose an elastic Graph Transformer framework (EmoGT) by integrating the GCN with Transformer, considering its potentials in time-series classification. Be-

† Corresponding author

This work was supported in part by grants from STI 2030-Major Projects+2022ZD0208500, the National Natural Science Foundation of China (No. 61976135), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), SJTU Global Strategic Partnership Fund, Shanghai Marine Equipment Foresight Technology Research Institute 2022 Fund (No. GC3270001/012), SJTU Global Strategic Partnership Fund (2021 SJTU-HKUST), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on August 22,2023 at 01:33:27 UTC from IEEE Xplore. Restrictions apply.

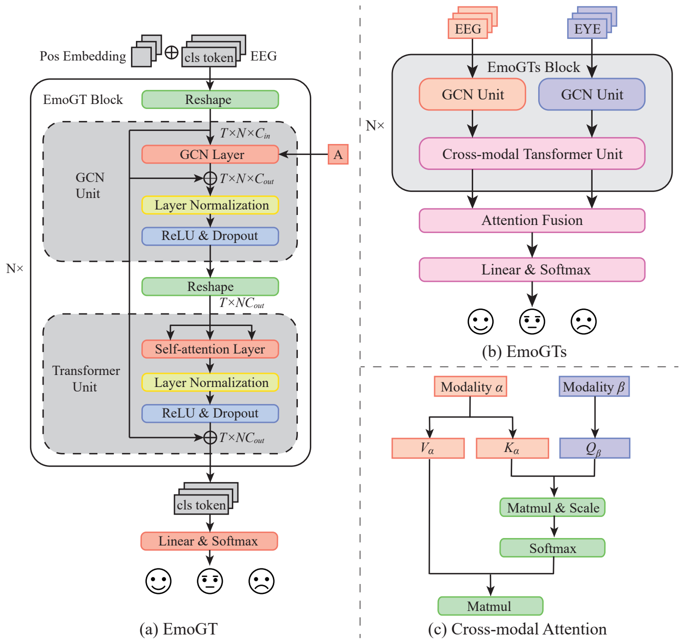


Fig. 1. (a) Illustration of EmoGT. (b) The simplified structure of EmoGTs. (c) The details of the Cross-modal Attention Layer in the Cross-modal Tansformer Unit.

sides, EmoGT can be used as individual building blocks for multimodal ER with two specially designed structures, which increase its robustness in real-world applications. Note that in computer vision, Zheng *et al.* [14] combined GCN and Transformer for whole slide image classification. They used GCN to propagate and aggregate information through image patches and used Transformer to learn the attention of different patches. Those two modules are operated in the same dimension (image patch-wise) which differs from our work.

2. METHODS

2.1. EmoGT

The overall architecture of EmoGT is shown in Figure 1(a). The backbone of EmoGT has N EmoGT blocks. Each block consists of two basic units, which are GCN Unit and Transformer Unit. The GCN Unit aims to fully learn the spatial relationship among electrodes while the Transformer Unit utilizes the temporal information of EEG.

GCN Unit A number of existing approaches have verified the effectiveness of GCN in EEG-based ER since GCN can model the electrodes on the scalp and catch the relationships of them by aggregating the information from their neighbors. We construct the EEG signals as a graph $G = \{V, E\}$, where V denotes the set of vertices and E denotes the set of edges. The adjacency matrix $A \in \mathbb{R}^{N\times N}$ represents E, which implies the connections between EEG channels, and N is the number of EEG channels i.e. $N = |V|$. The input EEG features is denoted by $X = (X_1, X_2, ..., X_T) \in \mathbb{R}^{T\times N\times C}$, where T is the number of samples in time series, and C is the dimension of each channel.

Kipf *et al.* [15] proposed the GCN model as follows

$$H_g^l = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^l W^l, \tag{1}$$

where $\hat{A} = A + I$, and I is the identity matrix. $\hat{D}$ denotes the degree matrix which is diagonal and $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$. H^l represents the input features of l-th layer while W^l is the weight matrix, $H^l \in \mathbb{R}^{T\times N\times C_{in}}$, $W^l \in \mathbb{R}^{C_{in}\times C_{out}}$. The output of l-th GCN layer is $H_g^l \in \mathbb{R}^{T\times N\times C_{out}}$. Note that $H^0 = X$.

The adjacency matrix A is essential for learning the graph representation as it describes the topology of EEG signals. We want the model to learn the topological structure from the data, so each entry of A_{ij} which shows the weight of connection between channel i and channel j is set learnable. Furthermore, A is defined as a symmetric matrix to avoid overfitting. For all GCN layers, we use the same shared adjacency matrix.

After the GCN layer, we add a skip connection from the input. The layer normalization is applied then and we choose ReLU as the activation function followed by a dropout layer.

Transformer Unit Transformer has achieved a great success in nature language processing, computer vison and other different areas. Before the Transformer Unit, the output of the GCN Unit $H_g^l \in \mathbb{R}^{T\times N\times C_{out}}$ is first reshaped to $H_t^l \in \mathbb{R}^{T\times NC_{out}}$. The self-attention layer is similar with the original multi-head attention proposed by [9]. The input feature H_t^l is transformed to queries Q_i, keys K_i and values V_i:

$$Q_i = H_t^l W_i^Q, K_i = H_t^l W_i^K, V_i = H_t^l W_i^V, \tag{2}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{NC_{out}\times d}$ and $Q_i, K_i, V_i \in \mathbb{R}^{T\times d}$. The self-attention can be calculated as

$$Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d}})V_i. \tag{3}$$

We employ h heads self-attention here, and each head can be denoted by $head_i = Attention(Q_i, K_i, V_i)$. The output of multi-head attention is

$$H_t^{l+1} = Concat(head_1, head_2, ..., head_h)W^O, \tag{4}$$

where $W^O \in \mathbb{R}^{dh\times NC_{out}}$ and $H_t^{l+1} \in \mathbb{R}^{T\times NC_{out}}$.

The same as the GCN Unit, there is a layer normalization after the self-attention layer followed by ReLU and dropout. The skip connection is adopted later.

Class Token and Position Embedding For better classification, we prepend a class token to the input. It is learnable and with the same shape as the input EEG feature. The role of the class token is similar with the one in BERT. At the end of all EmoGT blocks, we apply a linear layer and a softmax layer to the class token for classification. Since the input EEG features are a continuous time sequence, we add learnable position embeddings to the input sequence to enable the sequence to carry temporal information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on August 22,2023 at 01:33:27 UTC from IEEE Xplore. Restrictions apply.

2.2. EmoGTs

The main differences between EmoGT and EmoGTs are the Cross-modal Transformer Unit and Attention Fusion when the other modules remain the same, as illustrated in Figure 1(b). Besides EEG, the other modality we employ is eye movements. Similar to EEG, the input eye movement features can be denoted as $X' = (X'_1, X'_2, ..., X'_T) \in \mathbb{R}^{T \times N' \times 1}$, where T is the number of samples in time series which is the same as EEG sequence, and N' is the dimension of eye movement features. Eye movements are also modeled as a graph since there are dependencies among them, such as the fact that blinks affect the central programming of saccades [16].

Cross-modal Transformer Unit Inspired by the work in [17], we propose the Cross-modal Transformer Unit to fuse modalities so that the model is able to make use of the complementary properties of different modalities. The core of the Cross-modal Transformer Unit is that the self-attention layer is replaced by the cross-modal attention layer as depicted in Figure 1(c).

We describe the Cross-modal Attention Layer here. Considering two modalities α and β, we denote the input of cross-modal attention layer as $H_\alpha \in \mathbb{R}^{T \times C_\alpha}$ and $H_\beta \in \mathbb{R}^{T \times C_\beta}$, respectively. The queries, keys, values can be calculated by

$$Q_\beta = H_\beta W_\beta^Q, K_\alpha = H_\alpha W_\alpha^K, V_\alpha = H_\alpha W_\alpha^V, \tag{5}$$

where $W_\beta^Q \in \mathbb{R}^{C_\beta \times d}$, $W_\alpha^K \in \mathbb{R}^{C_\alpha \times d}$ and $W_\alpha^V \in \mathbb{R}^{C_\alpha \times d}$. The cross-modal attention from α to β is defined as

$$CMA_{\alpha:\beta}(H_\alpha, H_\beta) = softmax(\frac{Q_\beta K_\alpha^T}{\sqrt{d}})V_\alpha. \tag{6}$$

In the same way, the cross-modal attention from β to α can be calculated by $CMA_{\beta:\alpha}(H_\beta, H_\alpha)$. Different modalities interact with each other through this way to learn the complementary information. Note that we also employ multi-head attention in Cross-modal Transformer Unit.

Attention Fusion We use O_α and O_β to represent the class token of $CMA_{\alpha:\beta}$ and $CMA_{\beta:\alpha}$ which is the output of modality α and β after all EmoGTs Blocks, where $O_\alpha \in \mathbb{R}^{d_\alpha}, O_\beta \in \mathbb{R}^{d_\beta}$. We first transform O_α and O_β to the same dimension:

$$\hat{O}_\alpha = O_\alpha W_\alpha, \hat{O}_\beta = O_\beta W_\beta, \tag{7}$$

where $W_\alpha \in \mathbb{R}^{d_\alpha \times d}$ and $W_\beta \in \mathbb{R}^{d_\beta \times d}$. Then we compute the attention weights μ_α and μ_β by

$$\hat{\mu}_\alpha = \langle \hat{O}_\alpha, W^A \rangle, \hat{\mu}_\beta = \langle \hat{O}_\beta, W^A \rangle, \tag{8}$$

$$\mu_\alpha, \mu_\beta = softmax(\hat{\mu}_\alpha, \hat{\mu}_\beta), \tag{9}$$

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

$$O = \mu_\alpha \hat{O}_\alpha + \mu_\beta \hat{O}_\beta. \tag{10}$$

The above methods we talk about are using two modalities. It is worth mentioning that it can be easily extended to three or more modalities by adding extra streams of cross-modal attention between all pairs of modalities.

3. EXPERIMENTS

3.1. Datasets and Implementation Details

We test our models on three public datasets for ER, including SEED, SEED-IV, and SEED-V [18–20] , to thoroughly verify their performance. For EEG, differential entropy [21] is extracted within a nonoverlapping 1s time window from 5 frequency bands of every sample. For the eye movements, 33 features (e.g. pupil diameter, dispersion, fixation duration) are extracted as described in [7]. The input features are obtained by sliding an overlapping window with the size of $T = 5$ which is the same as the existing method [13] to make results more comparable. The batch size is 32 and the dropout rate is 0.5. The number of blocks is 4 and the learning rate is ranging from 3e-5 to 1e-3. The number of heads h and embedding dimension C_{out} are tuned from $\{2, 4\}$ and $\{16, 32, 64\}$, respectively. We choose cross entropy as the loss function and use Adam to optimize the parameters. For data division, we use the first 9 trials as training data and the remaining 6 trials as testing data in SEED. As for SEED-IV, the first 16 trials are training data while the remaining 8 trials are testing data. For SEED-V, however, we adopt 3-fold cross-validation.

3.2. Experimental Results

Comparison with Single Modal Methods To validate the efficiency of EmoGT, we compare it with other representative models, ranging from the baseline method DGCNN to

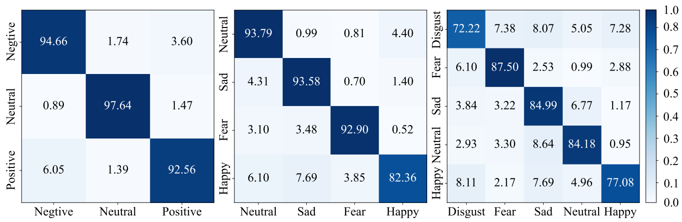
Table 1. Performances (Avg./Std.) on 3 datasets.

Method	SEED	SEED-IV	SEED-V
DGCNN [12]	90.40/08.49	69.88/16.29	-
BiDANN [22]	92.38/07.04	70.29/12.63	-
BiHDM [23]	93.12/06.06	74.35/14.09	-
RGNN [11]	94.24/05.95	79.37/10.54	-
MD-AGCN [13]	94.81/04.52	87.63/05.77	80.77/06.61
Transformer	92.35/07.55	86.40/11.45	77.32/08.19
G+LSTM	92.25/07.40	85.73/13.51	77.65/08.23
EmoGT	**95.02**/05.99	**91.20**/09.60	**82.73**/07.21
Fuzzy [24]	87.59/19.90	73.60/16.70	73.20/08.70
BDAE [20]	91.01/08.91	85.10/11.80	79.70/04.76
DCCA [20]	94.60/06.20	87.50/09.20	85.30/05.60
E-CMA	97.20/05.90	92.83/08.93	85.48/08.22
EmoGTs	**97.74**/05.43	**94.04**/08.69	**87.58**/08.62

* G+LSTM: replace Transformer Unit of EmoGT with LSTM.

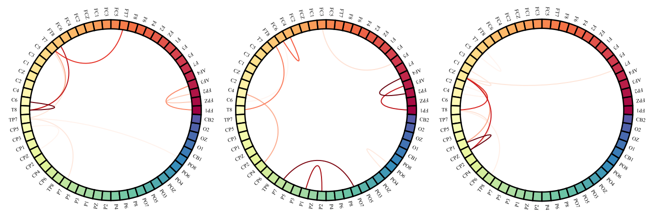
* E-CMA: EmoGTs without Cross-modal Attention Layer.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on August 22,2023 at 01:33:27 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c)

Fig. 2. Confusion matrices of EmoGT on three datasets. (a) SEED. (b) SEED-IV. (c) SEED-V. The vertical axis represents true labels; the horizontal axis predicted labels.



(a) SEED (b) SEED-IV (c) SEED-V

Fig. 3. Top 10 connections in the learned adjacency matrix. The darker the line, the higher the edge weight, and the stronger the connections between brain regions.

the state-of-the-art method MD-AGCN as reported in the first part of Table 1. In all, EmoGT outperforms all other models. For the existing best results gained by MD-AGCN, our model still improves by 1.5% on average for all three datasets. Noticeably, in SEED-IV, EmoGT greatly surpasses the current best result by about 3.5% and even reaches 91.2%, which is quite a high accuracy in four-labeled classification. The numbers undoubtedly demonstrate the powerful information extraction and processing ability of EmoGT.

The confusion matrices of EmoGT in Figure 2 depict how well it can distinguish between emotions on different datasets. For SEED and SEED-IV, our model is comparatively good at recognizing the neutral states and it does not behave well with positive emotions. Interestingly, in SEED-V, the disgust turns out to be the biggest obstacle for EmoGT to classify with only 72.22%. To sum up, EmoGT is likely to misclassify the positive emotions and the disgust emotions, which is of the same trend as other models.

Comparison with Multimodal Methods We pick up three most advanced methods in ER to compare with EmoGTs. The other modality we choose is eye movements, the combination of which with EEG signals has been proven to be a promising approach with high interpretability [24]. Results on three datasets are presented in the second part of Table 1. Though the latest best results achieved by DCCA are excellent, the EmoGTs still outshines them on all datasets. Remarkably, on the four-label classification task, our model has an amazing accuracy at 94% which surpasses the state-of-the-art performance by about 6.5%. Besides, EmoGTs also exceeds DCCA by 2% on the five-class task. The accuracy is even competitive in multi-class classification. Moreover, the decent standard deviations prove that the model has a relatively stable performance. In summary, EmoGTs represents the best level in multimodal ER, and these satisfying results inspire similar multimodal classification tasks with EmoGTs's solution.

Ablation Study As shown in italic lines in Table 1. The average performance gap reaches more than 5% with better stability. The effectiveness of extracting temporal information with Transformer has been proved by the comparison of LSTM-based variant G+LSTM and EmoGT. The latter behaves much better than the former with 3% to 5% higher accuracies. The gap between E-CMA and EmoGTs confirms that the cross-modal attention mechanism we introduce is very useful for multimodal learning, and it indeed employs the complementary properties of different modalities to achieve better performance.

Visualization of the Learned Spatial Connections In Figure 3, we present the brain connections when emotions elicit among EEG channels for all three datasets learned by EmoGT. The 10 strongest connections of SEED and SEED-IV are similar to each other while the one of SEED-V differs a lot. From the figure, we can see that the connections mainly concentrate at the temporal lobes on both sides of the brain together with the frontal lobe in SEED and SEED-IV, demonstrating by connections such as FT8 to T8 and FP1 to AF3, which is consistent with previous studies [18, 25]. However, the lines in the plot of the SEED-V illustrate the strong connections between the central areas and the temporal lobes, e.g. CZ to T8, as well as within the central areas of the brain like CP1 to CPZ and CZ to CPZ. One plausible explanation from cognitive neuroscience perspective is that because emotional processes with more emotion categories involve highly distributed neural circuits in the brain. As the number of emotions increases, more brain regions are activated to respond [26], resulting in global inter-channel connections.

4. CONCLUSIONS

In this paper, we have developed an elastic emotion recognition framework EmoGT to better exploit the EEG information from both temporal and spatial domain concurrently. It is achieved by integrating the GCN, for spatial relation extraction, with Transformer, for time-series information processing. With a cross-modal attention mechanism activated, EmoGT can be used as building block for individual modalities to tackle multimodal inputs. Our models outperform all state-of-the-art results by a large margin. The proposed framework sheds light on the EEG-based emotion recognition task by leveraging both spatial and temporal dynamics in EEG. Our future work will focus on generalization ability of EmoGT by adapting more diverse inputs.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on August 22,2023 at 01:33:27 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] C. Brunner et al., "BNCI Horizon 2020: towards a roadmap for the BCI community," *Brain-Computer Interfaces*, vol. 2, no. 1, pp. 1–10, 2015.

[2] G. L Ahern et al., "Differential lateralization for positive and negative emotion in the human brain: EEG spectral analysis," *Neuropsychologia*, vol. 23, no. 6, pp. 745–755, 1985.

[3] B. Burle et al., "Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view," *International Journal of Psychophysiology*, vol. 97, no. 3, pp. 210–220, 2015.

[4] R. Jenke et al., "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[5] V. J Lawhern et al., "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, pp. 056013, 2018.

[6] W. Zaremba et al., "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[7] J. Ma et al., "Emotion recognition using multimodal residual LSTM network," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 176–183.

[8] S. Alhagry et al., "Emotion recognition based on EEG using LSTM recurrent neural network," *Emotion*, vol. 8, no. 10, pp. 355–358, 2017.

[9] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] L. F. Barrett et al., "Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain," *Current Opinion in Neurobiology*, vol. 23, no. 3, pp. 361–372, 2013.

[11] P. Zhong et al., "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans on Affective Computing*, 2020.

[12] T. Song et al., "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2018.

[13] R. Li et al., "A Multi-domain adaptive graph convolutional network for EEG-based emotion recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5565–5573.

[14] Y. Zheng et al., "A graph-transformer for whole slide image classification," *arXiv preprint arXiv:2205.09671*, 2022.

[15] T. N Kipf et al., "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[16] K. G Rottach et al., "Properties of horizontal saccades accompanied by blinks," *Journal of Neurophysiology*, vol. 79, no. 6, pp. 2895–2902, 1998.

[17] Y. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of Association for Computational Linguistics*. NIH Public Access, 2019, vol. 2019, p. 6558.

[18] W. L. Zheng et al., "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[19] W. L. Zheng et al., "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Trans on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[20] W. Liu et al., "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans on Cognitive and Developmental Systems*, 2021.

[21] R. Duan et al., "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2013, pp. 81–84.

[22] Y. Li et al., "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Trans on Affective Computing*, vol. 12, no. 2, pp. 494–504, 2018.

[23] Y. Li et al., "A novel bi-hemispheric discrepancy model for eeg emotion recognition," *IEEE Trans on Cognitive and Developmental Systems*, vol. 13, no. 2, pp. 354–367, 2020.

[24] Y. Lu et al., "Combining eye movements and EEG to enhance emotion recognition," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015, pp. 1170–1176.

[25] X. Wu et al., "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition," *Journal of neural engineering*, vol. 19, no. 1, pp. 016012, 2022.

[26] T. Dalgleish, "The emotional brain," *Nature Reviews Neuroscience*, vol. 5, no. 7, pp. 583–589, 2004.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on August 22,2023 at 01:33:27 UTC from IEEE Xplore. Restrictions apply.