ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446937

FUNCTIONAL EMOTION TRANSFORMER FOR EEG-ASSISTED CROSS-MODAL EMOTION RECOGNITION

Wei-Bang Jiang[*], *Ziyi Li*[*], *Wei-Long Zheng, Bao-Liang Lu*[†]

Department of Computer Science and Engineering, Shanghai Jiao Tong University, China

ABSTRACT

Multimodal emotion recognition based on electroencephalography (EEG) and eye movements has attracted increasing attention due to their high performance and complementary properties. However, there are two challenges that hinder its practical applications: the inconvenient EEG data collection and high-cost data annotation. In contrast, eye movements are convenient to obtain and process in real scenarios. To combine high performance of EEG and easy setups of eye tracking, we propose a novel EEG-assisted Contrastive Learning Framework with a Functional Emotion Transformer (ECO-FET) for cross-modal emotion recognition. ECO-FET leverages both the functional brain connectivity and the spectral-spatial-temporal domain of EEG signals simultaneously, which dramatically benefit the learning of eye movements. The whole process consists of three phases: pre-training, test, and fine-tuning. ECO-FET exploits the complementary information provided by multiple modalities during pre-training in order to improve the performance of unimodal models. In the pre-training phase, unlabeled EEG and eye movement data are fed into the model to contrastively learn the emotional latent representations between the two modalities, while in the test phase, eye movements and few labeled EEG samples are used to predict different emotions. Experimental results on three public datasets demonstrate that ECO-FET surpasses the state-of-the-art dramatically.

Index Terms— EEG, eye movement, cross-modal emotion recognition

1. INTRODUCTION

EEG is considered reliable in emotion recognition tasks since it reflects the internal physiological responses and has strong objectivity [1]. Nevertheless, single modalities usually have limited generalization and robustness [2]. To improve performance, researchers have been working on multimodal emotion recognition, which takes advantage of complementary information across different modalities such as EEG and eye movements [3, 4]. They also found that EEG has a better ability to recognize emotions than eye movements [5]. However, there are two challenges in deploying such multimodal systems in real scenarios. The first one is the collection of EEG data, which takes time and efforts to set up due to the inconvenience of using EEG electrode caps. The other one is data annotation. Emotion is complex, subjective, and continuous state. Data annotation usually requires many efforts and sophisticated experimental designs.

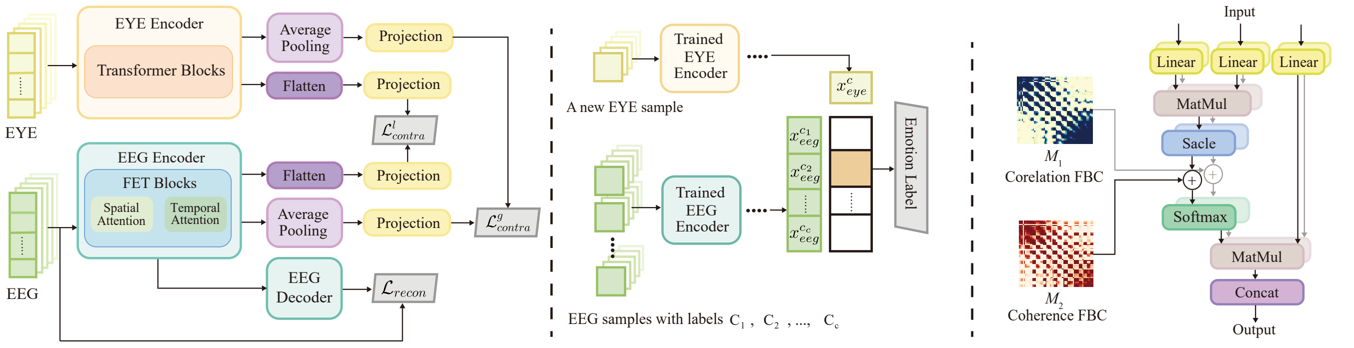
In contrast with EEG, eye movements are relatively easy to acquire in practical scenarios. The dependence on EEG can be reduced if we can figure out a method that utilizes EEG to help eye movements decode emotions during the training phase. Jiang *et al.* first applied a cross-modal transfer method, which regressed eye movements into multimodal features [6]. Yan *et al.* improved this approach by directly generating multimodal features from eye movements [7]. However, all of these approaches require accurate labels for cross-modal learning. In addition, since EEG signals are more complicated than eye movements, how to process the spectral-spatial-temporal domain of EEG signals effectively remains a challenging problem.

Recent progresses in contrastive learning [8, 9, 10] provide a new insight to deal with cross-modal learning. Inspired by these works, we propose a novel EEG-assisted cross-modal contrastive learning framework with a Functional Emotion Transformer which contrastively learns the emotional latent representations and transfers knowledge between EEG and eye movements in a self-supervised way. In the test phase, only eye movements are fed into the model to get the predicted emotions. To evaluate the effectiveness of our proposed ECO-FET, we conduct experiments on three public datasets. Experimental results demonstrate that our self-supervised approach achieves comparable performance with some supervised baselines. When we use the labeled data to fine-tune the pre-trained model, ECO-FET can drastically outperform the state-of-the-art. Therefore, our framework can leverage the high performance of multiple modalities and

[*] Equal contribution

[†] Corresponding author

This work was supported in part by grants from STI 2030-Major Projects+2022ZD0208500, National Natural Science Foundation of China (Grant No. 62376158), Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZD ZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.



(a) Pre-training process

(b) Test process

(c) Spatial attention mechanism in FET block

Fig. 1. Illustration of the pre-training, test processes and the spatial attention mechanism in the proposed ECO-FET.

easy setups of specific modalities when deployed.

2. METHODOLOGY

2.1. Overview

The proposed framework ECO-FET contains three processes: pre-training, test, and fine-tuning, as shown in Figure 1. The pre-training process is to minimize the distances of positive sample pairs while maximizing those of negative sample pairs. There can be more than one category of emotions stimulated during a video clip and even if there is only one type, the arousal of this emotion may vary over time. Therefore, only multimodal samples from the same time step have the same emotional state. Based on these considerations, we treat the EEG and eye movement pairs from the same time step as positive pairs while pairs in different time steps are negative. By this means the model is anticipated to excavate the emotion-related information that exists in both modalities.

2.2. Functional Emotion Transformer

Compared with eye movements, EEG signals have more complicated information and higher spatial-temporal resolution. Hence, a powerful backbone that can effectively decode emotions from EEG is urgently needed. We adapt the Functional Emotion Transformer (FET) and a critical subnetwork selection approach for self-supervised pre-training to incorporate functional brain connectivity (FBC) into FET [11, 12]. Note that for eye movements, we just use the vanilla Transformer as the backbone.

Functional Brain Connectivity We extract FBC M_1 and M_2 from the Pearson's correlation coefficient (temporal domain) and spectral coherence matrices (spectral domain) respectively, as shown in Figure 1(c) and Algorithm 1. We make two significant improvements. First, as ECO-FET is a self-supervised algorithm, labels are not allowed to use in the pre-training process. So, the correlation/coherence matrix is averaged over the same clip instead of the same label. Second, we take the degree of importance (the number of occurrences) of these connections into consideration.

Algorithm 1: The calculation of the functional brain connectivity (FBC)

```
Input: Preprocessed EEG signals $X_{pre}$, the number of clips in the training
       set $N_c$, the number of frequency band $N_f$ and threshold $t$.
Output: FBC $M$.
Get correlation/coherence matrix $X_{co}$ from $X_{pre}$;
$M \leftarrow zeros(62, 62)$;
for $i \leftarrow 1$ to $N_c$ do
    for $j \leftarrow 1$ to $N_f$ do
        Average the correlation or coherence matrices over the same clip:
          $A^{ij} \leftarrow mean(X^{ij}_{co})$;
        Sort the absolute values in the upper triangular position of $A^{ij}$:
          $A^{ij}_* \leftarrow sort(abs(triu(A^{ij})))$;
        Get the index of the first $t$ percentage of values in $A^{ij}_*$:
          $I^{ij} \leftarrow index(A^{ij}_*, t)$;
        $M[I^{ij}] \leftarrow M[I^{ij}] + 1$;
    end
end
$M \leftarrow normalize(M)$;
Return: $M$
```

Spatial and Temporal Attention We formulate the EEG input as $X \in \mathbb{R}^{N_e \times D_s}$, where N_e is the number of electrodes and D_s is the dimension of the spectral feature. The spatial attention layer applies self-attention mechanism among all the electrodes. If we solely apply the attention mechanism, the structural information between electrodes depends only on their semantic similarity while ignoring the structural prior itself. Thus, the correlation or coherence FBC M_i is added to attention attention matrix of different heads according to the following formula:

$$A_i = softmax(\frac{Q_i K_i^T}{\sqrt{D_{att}}} + M_i)V_i, \tag{1}$$

where $Q_i, K_i, K_i \in \mathbb{R}^{D_s \times D_{att}}$. Subsequently, the temporal attention layer is designed to make use of the continuous feature sequence for context learning.We denote $X^j_{eeg} \in \mathbb{R}^{S \times D_{eeg}}$ as the input sequence, where S is the time window size and $D_{eeg} = N_e D_s$ and j means the j-th sequence. The temporal attention is calculated as the standard self-attention [13].

2.3. The Self-supervised Pre-training Process

Encoders and the Contrastive Loss Assume that the batch size is N. We have N positive EEG and eye movement pairs and $N^2 - N$ negative pairs. Let $X_{eeg} \in \mathbb{R}^{N \times S \times D_{eeg}}$ represent the EEG input and $X_{eye} \in \mathbb{R}^{N \times S \times D_{eye}}$ denotes the eye movement input. X_{eeg} and X_{eye} are projected into a common embedding space through the following equation.

$$X_{eeg}^{g} = f_{P_{eeg}}(f_{ave}(f_{E_{eeg}}(X_{eeg}))), \tag{2}$$

where $X_{eeg}^{g} \in \mathbb{R}^{N \times D_f}$. $f_{E_{eeg}}$ and $f_{E_{eye}}$ are encoder s of EEG and eye movements respectively, which have l_{eeg} and l_{eye} blocks. f_{ave} is the average pooling layer which average the features across the time window dimension S. $f_{P_{eeg}}$ and $f_{P_{eye}}$ are the projection heads. Similarly, we can get X_{eye}^{g}. X_{eeg}^{g} and X_{eye}^{g} are conveyed to the InfoNCE loss [14] is utilized to get the global contextual contrastive loss $\mathcal{L}_{contra}^{g}$.

The global contextual contrastive loss focuses on optimizing the distances across the whole sequences of time windows. Nevertheless, it ignores the local features between individual samples, which differ in arousal even within the same time window. These inter-sample differences are beneficial for the emotional model learning. Therefore, we compute the local features as follow:

$$X_{eeg}^{l} = f_{P_{eeg}}(f_{fla}(f_{E_{eeg}}(X_{eeg}))), \tag{3}$$

where $X_{eeg}^{l} \in \mathbb{R}^{(NS) \times D_f}$. f_{fla} is the flatten layer to convert the dimensions of features to $(NS) \times D_{eeg}$ and $(NS) \times D_{eye}$. Similar as $\mathcal{L}_{contra}^{g}$, the local contrastive loss $\mathcal{L}_{contra}^{l}$ is computed using X_{eeg}^{l} and X_{eye}^{l}.

Decoder and the Reconstructed Loss It is believed that EEG data has more valuable emotion-related information [15], so we want the encoder of eye movements $f_{E_{eye}}$ to learn as much key information of EEG data as possible during pre-training. Based on this consideration, we add an EEG decoder $f_{D_{eeg}}$ and compute the reconstructed loss to restrict the EEG encoder $f_{E_{eeg}}$ to remain the core information of EEG data. Since we expect some beneficial changes in the encoder of eye movements, this restriction is not applied to the eye movement encoder. The decoder $f_{D_{eeg}}$ has similar structure with the encoders $f_{E_{eeg}}$ and $f_{E_{eye}}$, i.e., l_{eeg} FET blocks. The decoder output X_{eeg}^{o} is computed as

$$X_{eeg}^{o} = f_{D_{eeg}}(f_{E_{eeg}}(X_{eeg})). \tag{4}$$

Finally, the decoder output and the EEG initial input are utilized to calculate the MSE (Mean-Squared) loss:

$$\mathcal{L}_{recon} = \|X_{eeg}^{o} - X_{eeg}\|^2. \tag{5}$$

The total loss is the weighted sum of the above losses:

$$\mathcal{L} = \mathcal{L}_{contra}^{g} + \alpha \mathcal{L}_{contra}^{l} + \beta \mathcal{L}_{recon}, \tag{6}$$

where α and β are the trade-offs.

2.4. Test Process of the pre-training process

Figure 1(b) illustrates the test process in which only one new eye movement sample is taken as an example. Assume that the number of emotion labels is C. We randomly select C existing EEG samples with different labels $C_1, C_2, ..., C_c$. The new eye movement sample $X_{eye}^{t} \in \mathbb{R}^{1 \times D_{eye}}$ and the C selected EEG samples denoted as $X_{eeg}^{t} \in \mathbb{R}^{C \times D_{eeg}}$ are then passed through the same pre-trained model introduced above. The calculation of embedding features $X_{eeg}^{c} \in \mathbb{R}^{C \times D_f}$ is similar as equation 3 but the input is X_{eeg}^{t}. $X_{eye}^{c} \in \mathbb{R}^{1 \times D_f}$ is obtained similarly.

The cosine similarity between the embedding of the new eye movement sample X_{eye}^{c} and those of EEG samples X_{eeg}^{c} is then calculated. The greater the similarity, the more likely the categories of these two kinds of samples from different modalities are the same. Therefore, we assign the label of EEG sample which achieves the largest similarity value to this new eye movement sample.

2.5. Fine-tuning Process

The fine-tuning process simplifies the emotion recognition procedure by discarding EEG, i.e., we only utilize the eye movement data as input. The fine-tuning process follow the conventional supervised learning paradigms. we employ the trained eye movement encoder and a classifier, i.e., l' transformer blocks with a full-connected layer and a softmax activation function. In order to retain the beneficial information learned from EEG in the contrastive pre-training, we freeze the encoder and only tune the classification block. The cross-entropy loss is employed.

3. EXPERIMENTS

3.1. Datasets and Implementation Details

We comprehensively evaluate the performance of our framework on a series of three public multimodal emotion recognition datasets, SEED [16], SEED-IV [17], and SEED-V [18] with 3, 4 and 5 emotions, respectively. For EEG, we use the 310-dimension differential entropy (DE) feature as in [16, 19] and for eye movement, 33 features are calculated as in [17]. The data division is the same for all the employed methods as in [7].

The threshold t is set to 0.2. The number of FET blocks l_{eeg} and l_{eye} is 2 and 3, respectively, while l' in the classification block is set to 1. The dimension D_{att} and the hidden dimension in feed forward block are tuned from the combination of $\{(128, 32), (64, 16), (128, 16)\}$, where the former number in the parentheses represents the dimension of EEG and the latter represents that of eye movement. For the projection head in the pre-training process, the dimension of the

Table 1. The accuracies (Avg./Std.) of different methods on three datasets. PT denotes pre-train and FT denotes fine-tune.

Method	SEED	SEED-IV	SEED-V
BDAE-regressor [6]	75.72/08.87	73.49/07.02	72.80/05.07
BDAE-cGAN [7]	81.02/08.04	75.74/06.66	73.66/06.05
ECO-Transformer (PT)	77.03/09.80	66.71/10.53	57.50/04.19
ECO-Transformer (FT)	90.39/10.38	82.41/10.80	76.80/05.79
ECO-FET (PT)	83.35/07.98	69.57/14.02	59.48/04.05
ECO-FET (FT)	**93.69**/08.22	**87.76**/09.19	**77.13**/04.16
w/o FBC (PT)	82.61/08.38	69.22/13.54	58.84/04.68
w/o FBC (FT)	92.32/09.36	86.71/09.67	76.95/05.19
w/o Decoder (PT)	80.57/09.70	67.99/12.63	58.02/04.05
w/o Decoder (FT)	93.20/09.18	87.04/09.56	74.71/04.63
w/o local (PT)	81.91/08.74	68.19/12.96	57.60/03.52
w/o local (FT)	91.53/10.52	86.66/09.41	75.62/06.44
Scratch	86.06/10.99	78.16/12.00	71.74/05.94

common embedding space D_f is set to 16. The learning rate is set to 0.0001 or 0.001 while the batch size is 32. The coefficients α and β are tuned from set $\{0.1, 0.01\}$. The Adam optimizer is applied.

3.2. Experimental Results

To validate the performance of ECO-FET, we consider several typical cross-modal methods. Average accuracies and standard deviations are listed in Table 1. Note that BDAE-cGAN is the state-of-the-art cross-modal method. We replace FET with Transformer as the EEG backbone and discard the local contrastive loss to testify the performance of pure contrastive learning framework, denoted as ECO-Transformer.

As shown in Table 1, without any information from emotion labels in the training process, our pre-trained model achieves comparable performance compared with the single-modal supervised baseline method using eye movement signals in the first line. ECO-FET even surpasses the supervised baseline on SEED. When labeled data are available, ECO-FET can be easily fine-tuned to obtain more excellent classification accuracies, as suggested in the last row of Table 1. Our model has a huge advantage of over 12%, 12%, and 3% on SEED, SEED-IV, and SEED-V, compared with BDAE-cGAN. The quantitative comparison reflects the superiority of our cross-modal framework in decoding emotions from eye movements. Compared with ECO-Transformer, ECO-FET has a better performance in both pre-training and fine-tuning. The interesting results indicate that a powerful EEG feature extractor can also facilitate the learning of eye movements.

Effectiveness of Pre-training We train the whole encoder of eye movements and the classifier from scratch to compare performance with the fine-tuned model. The results are presented in Table 1. The accuracies are enhanced by approximately 6% to 9% on the three datasets. Meanwhile, the smaller standard deviations also substantiate that the fine-tuned model is more robust and stable. Therefore, we can deduce that the contrastive learning framework does enable eye movements to extract emotional latent representations with the help of EEG at training time.

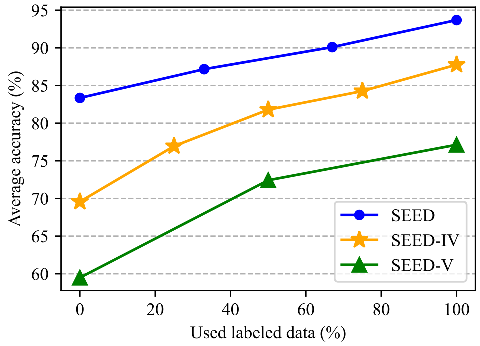


Fig. 2. Fine-tuning on different amount of labeled data.

Ablation Study We conduct ablation study to verify the effectiveness of different components of ECO-FET. Results are presented in Table 1. We remove the functional brain connectivity from FET and observe a decrease in accuracies, which demonstrates the importance of exploring functional brain connectivity in our model. In the pre-training process, we add a decoder for EEG signals aiming to keep the EEG information through the MSE loss. With the decoder, the accuracies of the pre-training model with the decoder are increased for all datasets, which demonstrates its effectiveness. Besides, we experimentally verify whether the local contrastive loss works. When we do not employ the local contrastive loss, the accuracies drop by 1% to 2%. These results prove the effectiveness of the local contrastive loss.

Fine-tuning with Partial Labeled Data We investigate the performance of our ECO-FET when there are fewer labeled data since it's common that in real life labeled data are limited. Experimental results are depicted in Figure 2. Note that we utilize different numbers of video clips such that each proportion contains all categories of emotions. The accuracies rise as expected when used labeled data get more and more. Compared with BDAE-cGAN, ECO-FET gets comparable accuracies only using about 50% labeled training data on SEED-V. Furthermore, our framework surpasses BDAE-cGAN with 25% labeled data on SEED-IV, and even acquires a better performance without any labeled data on SEED, which brings emotion recognition a step closer to practical application.

4. CONCLUSIONS

In this paper, we propose a self-supervised method called ECO-FET based on contrastive learning for cross-modal emotion recognition. This model offers a potential solution to improve the emotion classification performance of an easy setup but weaker modality by exploiting a stronger modality.

5. REFERENCES

[1] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu, "Emotional state classification from eeg data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.

[2] Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomedical Engineering*, vol. 6, pp. 1346 – 1352, 2022.

[3] Mohammad Soleymani, Maja Pantic, and Thierry Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2011.

[4] Wei-Long Zheng, B. Dong, and Bao-Liang Lu, "Multimodal emotion recognition using EEG and eye tracking data," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5040–5043, 2014.

[5] 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*. IEEE, 2014, pp. 5040–5043.

[6] Huangfei Jiang, Xiya Guan, Wei-Ye Zhao, Li-Ming Zhao, and Bao-Liang Lu, "Generating multimodal features for emotion classification from eye movement signals.," *Australian Journal of Intelligent Information Processing System*, vol. 15, no. 3, pp. 59–66, 2019.

[7] Xu Yan, Li-Ming Zhao, and Bao-Liang Lu, "Simplifying multimodal emotion recognition with single eye movement modality," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1057–1063.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[9] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[10] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik, "Self-supervised contrastive pretraining for time series via time-frequency consistency," *ArXiv*, vol. abs/2206.08496, 2022.

[11] 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*, vol. 19, no. 1, pp. 016012, jan 2022.

[12] Rui Li, Yiting Wang, and Bao-Liang Lu, "A multidomain adaptive graph convolutional network for eeg-based emotion recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5565–5573.

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

[14] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv e-prints*, pp. arXiv–1807, 2018.

[15] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu, "Combining eye movements and EEG to enhance emotion recognition," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[16] 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*, vol. 7, no. 3, pp. 162–175, 2015.

[17] 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*, vol. 49, no. 3, pp. 1110–1122, 2018.

[18] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[19] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 10, pp. 417–429, 2016.