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/ICASSP48485.2024.10447194

MULTIMODAL MULTI-VIEW SPECTRAL-SPATIAL-TEMPORAL MASKED AUTOENCODER FOR SELF-SUPERVISED EMOTION RECOGNITION

*Pengxuan Gao, Tianyu Liu, Jia-Wen Liu, Bao-Liang Lu, Wei-Long Zheng**

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

ABSTRACT

Emotion recognition is a primary and complex task in emotional intelligence. Due to the complexity of human emotions, utilizing multimodal fusion methods can enhance the performance by leveraging the complementary properties of different modalities. In this paper, we propose a Multimodal Multi-view Spectral-Spatial-Temporal Masked Autoencoder (Multimodal MV-SSTMA) with self-supervised learning to investigate multimodal emotion recognition based on electroencephalogram (EEG) and eye movement signals. Our experimental process comprises three stages: 1) In the pre-training stage, we employ MV-SSTMA to train feature extractors for EEG and eye movement signals; 2) In the fine-tuning stage, the labeled data are input to the feature extractors to fuse and fine-tune the features; 3) In the testing stage, our model is applied to recognize emotions with test data to calculate the accuracies of different methods. Our experimental results demonstrate that the multimodal fusion model outperforms the unimodal model on both SEED-IV and SEED-V datasets. In addition, the proposed model can still effectively recognize emotions with various ratios of missing data. These results underscore the efficiency of multimodal self-supervised learning and data fusion in emotion recognition.

***Index Terms*—** Emotion Recognition, Self-supervised Learning, EEG, Eye Movement, Multimodal Fusion

1. INTRODUCTION

Emotion recognition is a hot topic in current affective computing research. Traditional emotion recognition relies on non-physiological signals, such as facial expressions, body movements, and voice [1, 2]. However, these signals can be influenced by subjective factors among different individuals. In contrast, physiological signals like electroencephalogram (EEG) and eye movement (EYE) signals offer a more objective and precise depiction of emotions, providing a substantial advantage [3, 4, 5].

Since the scale of unlabeled physiological signals collected in the future is anticipated to expand progressively, and meanwhile, physiological signals are susceptible to noise interference in practical settings, it is neccessary to apply self-supervised learning, which aims to train models from data without explicit annotations or labels. Traditionally, most EEG-based self-supervised learning focused on contrastive methods. For instance, van den Oord *et al.* came up with Contrastive Predictive Coding (CPC) to use the temporal structure of sequential data to learn representations [6]. Chen *et al.* proposed SimCLR (Simple Framework for Contrastive Learning of Visual Representations), which learns representations by maximizing consistency between differently enhanced views of the same image [7]. For EEG data, Kostas *et al.* used contrastive self-supervised DNNs for pre-training and downstream fine-tuning tasks to learn the features of EEG signals [8]. Banville *et al.* employed self-supervised learning to learn representations of EEG signals based on temporal context prediction and contrastive predictive coding [9]. Moreover, Li *et al.* proposed a Multi-view Spectral-Spatial-Temporal Masked Autoencoder (MV-SSTMA) based on MAE [10]. MV-SSTMA incorporates a hybrid block structure of CNN and Transformer in its encoder and decoder.

Multimodal fusion integrates information from diverse sources to enhance model performance. In the past, early fusion techniques mainly relied on straightforward concatenation or weighting schemes. However, recent advancements in deep learning have explored more approaches to effectively combining various modalities. Deep learning models can extract high-level representations from raw data and capture complex intermodal relationships, yielding remarkable outcomes across various domains, including emotion recognition [11, 12, 13].

This paper makes the following contributions: (1) We propose a Multimodal Multi-view Spectral-Spatial-Temporal Masked Autoencoder (MMV-SSTMA) with self-supervised learning to learn robust EEG and eye movement representations. Through self-supervised learning, we avoid the large-

*Corresponding author (weilong@sjtu.edu.cn).

This work was supported in part by grants from National Natural Science Foundation of China (Grant No. 62376158), STI 2030-Major Projects+2022ZD0208500, 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.

scale labeling task. This approach maximizes the utilization of unlabeled data, which provides a practical method for scaling up multimodal emotion recognition training. (2) We generalize the proposed method from unimodal to multimodal settings. It not only validates the feasibility and rationality of emotion recognition using both EEG and eye movement signals, but also explores the complementary nature of both two modalities. (3) Extensive experiments on two public emotion datasets demonstrate the efficiency of the proposed model with multimodal pre-training even with missing data and only few labeled samples to calibrate.

2. METHODOLOGY

2.1. Base Model

Multi-view Spectral-Spatial-Temporal Masked Autoencoder (MV-SSTMA) is introduced as a solution to the challenges of decoding emotions from a limited set of labeled and damaged EEG data. MV-SSTMA comprises several components, including a spectral embedding layer, a spatial positional encoding layer, L hybrid encoders, and L symmetric hybrid decoders. The function of the encoders is to extract relevant features from the original EEG signals, while the decoders are designed to reconstruct the masked EEG data based on these extracted features.

Let the differential entropy (DE) feature of the original EEG signals be $X = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^{N\times C\times F}$, where N, C, F denote the number of samples in time series, the number of EEG channels and the set of frequency bands, respectively. By an overlapping window of T seconds, X is transformed into $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n) \in \mathbb{R}^{N\times C\times T\times F}$. Subsequently, in the spectral embedding layer, each sample $\tilde{x}_i$ is projected into a D-dimensional space via a linear layer.

In the spatial positional encoding layer, embedded sample is divided into patches in the dimension C, where each patch corresponds to one EEG channel. Sine-cosine positional encoding is added across the C dimension to retain position information. Meanwhile, random masking is applied to the data, allowing only a visible subset $\tilde{x}_i^v \in \mathbb{R}^{C_v\times T\times D}$ to be utilized as input for training.

The MV-SSTMA architecture incorporates L hybrid encoders, each consisting of a temporal multi-scale causal convolution layer and a spatial multi-head self-attention layer. In the temporal multi-scale causal convolution layer, three branches of causal convolution layers with long ($k_l \times 1$), medium ($k_m \times 1$) and short ($k_s \times 1$) kernel sizes are implemented to capture data characteristics in the temporal dimension. Subsequently, the output from the convolutional layer is fed into the multi-head self-attention layer. Assuming the input of this layer is $B_{in} \in \mathbb{R}^{C_v\times T\times D}$, the formulations for the convolutional layer and self-attention layer results are as follows:

$$B_s^l = \text{Attention}\left(\text{BN}\left(\text{CausalConvT}\left(B_{in}, (k_l, 1)\right)\right)\right), \quad (1)$$

$$B_s^m = \text{Attention}\left(\text{BN}\left(\text{CausalConvT}\left(B_{in}, (k_m, 1)\right)\right)\right), \quad (2)$$

$$B_s^s = \text{Attention}\left(\text{BN}\left(\text{CausalConvT}\left(B_{in}, (k_s, 1)\right)\right)\right), \quad (3)$$

where BN is the batch normalization operation, Attention is the Multi-head self-Attention operation, and B_s^l, B_s^m and B_s^s all have the same shape of $C_v \times T \times D$ as B_{in}.

The spatial brain summary $\tilde{B}^{(C_v)}$ is the sum of them:

$$\tilde{B}^{(C_v)} = B_s^l + B_s^m + B_s^s. \quad (4)$$

Following the spatial attention layer, $\tilde{B}^{(C_v)}$ is input to the normalization layer and the feed-forward network to obtain the data that are reconstructed in the decoder.

During the pre-training process, the symmetric decoder for EEG signals reconstruction is added, which comprises L CNN-Transformer hybrid blocks and a linear layer. It predicts values for each masked EEG channel.

In the testing process, we employ the feature extractor E to extract features from EEG signals, input them into the classifier, fine-tune it with a subset of labeled data, and evaluate its effectiveness on the test dataset.

2.2. Multimodal Model

Based on MV-SSTMA, our objective is to utilize multimodal fusion techniques to integrate EEG and eye movement data in order to enhance the performance of the original MV-SSTMA model. The experimental procedure is depicted in **Fig. 1**.

Pre-training In the pre-training stage, we leverage the structural and parametric similarities between EEG and eye movement signals. Two feature extractors are trained for these two signals, separately, utilizing nearly identical parameters and architectures. However, in the embedding layer, eye movement signals are mapped to the same dimensional space as EEG signals to maintain consistency in feature dimensions.

Fine-tuning With the learned encoder, we extract features from EEG and eye movement signals, respectively. Suppose the two feature vectors are denoted as $X^1 = \left[x_1^1, x_2^1, \cdots, x_n^1\right] \in \mathbb{R}^n$ and $X^2 = \left[x_1^2, x_2^2, \cdots, x_m^2\right] \in \mathbb{R}^m$, we can fuse them using Concatenation Fusion [14]:

$$X_{fusion} = Concat\left(\left[X^1, X^2\right]\right), \quad (5)$$

or Attention Mechanism Fusion [15]:

$$F = Concat\left(X^1, X^2\right), \quad (6)$$

$$\alpha = softmax\left(w^T \tanh\left(F\right)\right), \quad (7)$$

$$X_{fusion} = F\alpha^T, \quad (8)$$

where w is the linear layer parameter and α denotes the attention weight matrix. Meanwhile, Deep Canonical Correlation Analysis (DCCA) is employed to align the features and improve the correlation between different modalities [16]. Eventually, the fused features are input to a linear layer for classification, and then the classification loss and the correlation loss of DCCA are utilized together for optimization. This process

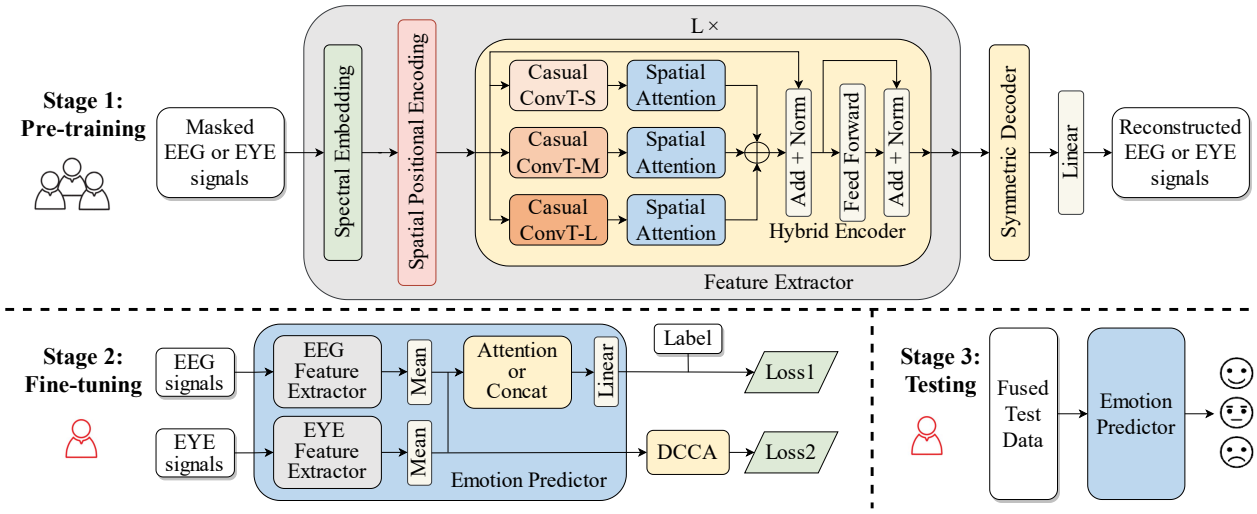

Fig. 1. The overall process of Multimodal Multi-view Spectral-Spatial-Temporal Masked Autoencoder.

is carried out using a few labeled data, which can either encompass the entirety or a subset of the data, aiming to derive a distinct classifier E^i for each subject i.

Testing The classifier E^i is utilized to test subject i and calculate the accuracy for each $i \in \{1, 2, \cdots, H\}$, and the average accuracy and standard deviation of all subjects are computed. These statistical measures are used to assess and compare the training effectiveness of different models.

3. EXPERIMENTAL SETTINGS

3.1. Datasets

In our experiment, we utilize two public emotion datasets.

SEED-IV dataset The SEED-IV dataset contains EEG and eye movement signals of 15 subjects (7 male and 8 female) of four different emotion states: happy, sad, fear and neutral emotions [17]. The 15 subjects participated in three sessions on different days with 24 trials each. For each session, the first 16 trials are set to be training data and the remaining 8 trials are set to be test data.

SEED-V dataset The SEED-V dataset contains EEG and eye movement signals for five emotion states: happy, sad, neutral, fear and disgust [18]. In one session, 6 male and 10 female subjects watched 15 movie clips (3 clips for each emotion) and each subject participated in the experiment for three sessions. The first 9 trials are used as training data and the remaining 6 trials are used as test data for each session.

3.2. Parameter Settings

To address the large numerical differences between EEG and eye movement features in the original datasets, we normalize the values of eye movement features into the range of 10 to 100 before we input them to the encoder, as part of the implementation of mixed-precision training.

Table 1. The average accuracy (%) and standard deviation (%) of EEG and eye movement signals with and without pre-training parameters on SEED-IV and SEED-V datasets.

Signal	Pre-training Parameters	Dataset			
		SEED-IV		SEED-V	
		Avg.	Std.	Avg.	Std.
EEG	w/o	56.97	9.56	71.29	10.76
	w	**85.40**	**5.25**	**79.47**	**9.50**
EYE	w/o	64.98	14.60	53.88	6.84
	w	**78.16**	**8.26**	**64.19**	**6.44**

In our experiment, the number of EEG channels $C = 62$, the set of frequency bands $F = 5$ (δ: 1-4 Hz, θ: 4-8 Hz, α: 8-14 Hz, β: 14-31 Hz, γ: 31-50 Hz), the size of the overlapping window $T = 10$, the spectral embedding size $D = 16$, the number of the hybrid block $L = 6$, and the multi-head dimension $H = 6$. The mask ratio in the spatial positional encoding layer is set to 30% by default. The learning rate is equal to $1.5e^{-4}$ in the pre-training stage and 0.001 in other stages. The hyperparameter of DCCA loss is 0.1.

4. EXPERIMENTAL RESULTS

4.1. Results for Pre-training

We compare the average accuracies and standard deviations under two conditions: with and without loading the pre-training parameters. The results are presented in **Table 1**. **Table 1** demonstrates that using pre-training parameters results in significantly improved performance for both EEG and eye movement signals compared to training without them. The higher average accuracy validates the effectiveness of pre-training parameters, while the lower standard deviation suggests improved model stability.

Table 2. The average accuracy (%) and standard deviation (%) of different models using all labeled data to fine-tune on SEED-IV and SEED-V datasets.

Model	Dataset			
	SEED-IV		SEED-V	
	Avg.	Std.	Avg.	Std.
MV-SSTMA-EEG	87.78	6.90	80.41	10.51
MAX	60.00	17.10	73.20	9.30
Fuzzy Integral	73.60	16.70	73.20	8.70
BDAE	85.10	11.80	79.70	**4.80**
MMV-SSTMA-Concat	89.56	6.19	81.03	9.31
MMV-SSTMA-Att	**90.16**	**5.44**	**81.38**	8.76

Table 3. The average accuracy (%) and standard deviation (%) of different models using partial labeled data to fine-tune on SEED-IV and SEED-V datasets.

Model	Dataset			
	SEED-IV		SEED-V	
	Avg.	Std.	Avg.	Std.
MV-SSTMA-EEG	87.03	**5.88**	79.07	10.95
MMV-SSTMA-Concat	**88.64**	6.71	79.13	8.72
MMV-SSTMA-Att	88.04	6.57	**79.20**	**8.49**

4.2. Results for Fine-tuning

Table 2 presents an overview of the outcomes during the fine-tuning stage when all labeled data are employed. In **Table 2**, MV-SSTMA-EEG is the model that only utilizes EEG signals, MAX, Fuzzy Integral and BDAE are the models coming from [19], MMV-SSTMA-Concat denotes the multimodal MV-SSTMA model that fuses EEG and eye movement features by concatenation, and MMV-SSTMA-Att represents the multimodal MV-SSTMA model that leverages an attention mechanism for the fusion of the two modalities. As demonstrated in **Table 2**, the models employing multimodal fusion methods exhibit higher average accuracies and lower standard deviations compared to the models utilizing EEG signals on both SEED-IV and SEED-V datasets, showing the potential of multimodal fusion to enhance emotion recognition performance and stability.

In particular, MMV-SSTMA-Att achieves the highest average accuracy of 90.16% on SEED-IV dataset and 81.38% on SEED-V dataset. This superiority may be attributed to the attention mechanism fusion, which offers a more rational integration of features across different modalities, assigning varying weights to different modal features to identify the valuable ones.

Furthermore, partial labeled training data are utilized to fine-tune the model. In this paper, 10 labeled data from the beginning of the training dataset are used for calibration. The results are presented in **Table 3**. Notably, the two models employing multimodal fusion methods consistently outperform the baseline. On SEED-IV dataset, MMV-SSTMA-Concat achieves the highest average accuracy of 88.64% and MV-SSTMA-EEG obtains the lowest standard deviation of 5.88%. Conversely, on the SEED-V dataset, MMV-SSTMA-Att obtains the highest average accuracy of 79.20% and the lowest standard deviation of 8.49%.

Table 4. The average accuracy (%) of different models using partial data with different mask ratios to fine-tune on SEED-IV and SEED-V datasets.

Model	Dataset					
	SEED-IV			SEED-V		
	30%	50%	70%	30%	50%	70%
MV-SSTMA-EEG	87.03	86.84	84.25	79.07	78.44	75.59
MMV-SSTMA-Concat	**88.64**	86.60	85.00	79.13	**84.15**	77.81
MMV-SSTMA-Att	88.04	**89.95**	**86.74**	**79.20**	81.06	**78.29**

Compared with the results in **Table 2**, **Table 3** reflects that the reduction in the amount of data leads to a decrease in the performance of each model. However, the multimodal fusion models with self-supervised learning consistently outperform the baseline, which demonstrates their potential to reduce the need for labeled data in various real-world applications.

4.3. Results for different mask ratios

We present the results of various mask ratios applied with partial labeled data on SEED-IV and SEED-V datasets in **Table 4**. The training data and the test data are masked using the same mask ratios. The findings indicate that for MV-SSTMA-EEG, the optimal mask ratio is 30%, resulting in the highest average accuracy (87.03% on SEED-IV, 79.07% on SEED-V) and for MMV-SSTMA-Concat, it is best to set the mask ratio to 30% on SEED-IV (88.64% average accuracy) and 50% on SEED-V (84.15% average accuracy). Conversely, for MMV-SSTMA-Att, setting the mask ratio to 50% yields the best performance with the highest average accuracy (89.95% on SEED-IV, 81.06% on SEED-V). Furthermore, it's worth noting that the multimodal models still outperform the baseline across different mask ratios.

5. CONCLUSIONS

In this paper, we develop the multimodal MV-SSTMA model to efficiently learn representations from EEG and eye movement signals with masked autoencoders as self-supervised learning. By employing fusion techniques based on concatenation and attention mechanism, we perform the modality fusion and model fine-tuning with a few labeled data. This approach yields better results than baseline models, demonstrating the feasibility of pre-training and multimodal fusion strategies, highlighting the complementary nature of EEG and eye movement signals. By using self-supervised learning and limited data fine-tuning, we reduce the dependence on labeled data and leverage unlabeled data.

6. REFERENCES

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] Byoung Chul Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, pp. 401, 2018.

[3] 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.

[4] 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, no. 3, pp. 417–429, 2017.

[5] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster, "Eye movement analysis for activity recognition using electrooculography," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 741–753, 2010.

[6] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *ArXiv Preprint ArXiv:1807.03748*, 2018.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[8] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz, "Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers in Human Neuroscience*, vol. 15, pp. 653659, 2021.

[9] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort, "Uncovering the structure of clinical EEG signals with self-supervised learning," *Journal of Neural Engineering*, vol. 18, no. 4, pp. 046020, 2021.

[10] 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*, 2022, pp. 6–14.

[11] Dana Lahat, Tülay Adali, and Christian Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[12] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu, "Emotion recognition using multimodal deep learning," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 521–529.

[13] Zhongmin Wang, Xiaoxiao Zhou, Wenlang Wang, and Chen Liang, "Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 4, pp. 923–934, 2020.

[14] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 284–288.

[15] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 207–212.

[16] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1247–1255.

[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] Tian-Hao Li, Wei Liu, Wei-Long Zheng, and Bao-Liang Lu, "Classification of five emotions from EEG and eye movement signals: Discrimination ability and stability over time," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 607–610.

[19] 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*, vol. 14, no. 2, pp. 715–729, 2021.