Emotion Recognition of Subjects With Hearing Impairment Based on Fusion of Facial Expression and EEG Topographic Map

Dahua Li, Jiayin Liu, Yi Yang, Fazheng Hou, Haotian Song, Yu Song, *Member, IEEE*, Qiang Gao, and Zemin Mao

***Abstract*— Emotion analysis has been employed in many fields such as human-computer interaction, rehabilitation, and neuroscience. But most emotion analysis methods mainly focus on healthy controls or depression patients. This paper aims to classify the emotional expressions in individuals with hearing impairment based on EEG signals and facial expressions. Two kinds of signals were collected simultaneously when the subjects watched affective video clips, and we labeled the video clips with discrete emotional states (fear, happiness, calmness, and sadness). We extracted the differential entropy (DE) features based on EEG signals and converted DE features into EEG topographic maps (ETM). Next, the ETM and facial expressions were fused by the multichannel fusion method. Finally, *a* deep learning classifier CBAM_ResNet34 combined Residual Network (ResNet) and Convolutional Block Attention Module (CBAM) was used for subject-dependent emotion classification. The results show that the average classification accuracy of four emotions recognition after multimodal fusion achieves 78.32%, which is higher than 67.90% for facial expressions and 69.43% for EEG signals. Moreover, visualization by the Gradient-weighted Class Activation Mapping (Grad-CAM) of ETM showed that the prefrontal, temporal and occipital lobes were the brain regions closely related to emotional changes in individuals with hearing impairment.**

Manuscript received 9 May 2022; revised 3 November 2022; accepted 27 November 2022. Date of publication 1 December 2022; date of current version 1 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62103299 and in part by the 2021 Tianjin Postgraduate Research and Innovation Project under Grant 2021YJSS092. *(Dahua Li and Jiayin Liu are co-first authors.) (Corresponding authors: Yu Song; Qiang Gao.)*

Dahua Li, Jiayin Liu, Fazheng Hou, Haotian Song, and Yu Song are with the Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems, School of Electrical Engineering and Automation, Tianjin University of Technology, Tianjin 300384, China (e-mail: lidah2005@163.com; ljysweet@yeah.net; hfzbrightsome@hotmail.com; yarafisong123@163.com; jasonsongrain@hotmail.com).

Yi Yang is with the Department of Electrical and Computer Engineering, Faculty of Science and Technology, University of Macau, Zhuhai, Macau (e-mail: yyflying@yeah.net).

Qiang Gao is with the Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems, TUT Maritime College, Tianjin University of Technology, Tianjin 300384, China (e-mail: gaoqiang@tjut.edu.cn).

Zemin Mao is with the Technical College for the Deaf, Tianjin University of Technology, Tianjin 300384, China (e-mail: maozemin@email.tjut.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3225948

***Index Terms*— Emotion recognition, facial expression, electroencephalogram topographic map, individuals with hearing impairment.**

I. Introduction

BRAIN-COMPUTER interface (BCI) technology is a method of recording neural signals such as Electroencephalogram (EEG) from the brain through the external devices. With the development of non-invasive signal acquisition techniques, BCI technology has made great contributions in many fields, including human-robot interaction, rehabilitation and neuroscience. In the field of neuroscience, BCI technology can classify the degree of mental disorders in patients with depression [1], and detect the degree of mental fatigue of drivers to prevent traffic accidents [2]. Affective computing is an important branch of BCI field, emotion is a complex psychological state or process of people. As a subjective experience of the human body, emotion can directly affect various aspects such as individual survival and development, interpersonal communication, and mental health.

Psychologists carried out the discrete model and dimensional model to describe the emotional states. Ekman et al. [3] proposed the concept of "basic emotions" as fear, anger, surprise, disgust, joy, sadness. The dimensional emotional model [4] turns the change of emotion into two-dimension, which represents emotional states in valence-arousal. Valence ranges from negative to positive, and arousal ranges from passive to active. Since then, many researchers have started to explore the field of affective computing based on discrete emotional model or dimensional emotional model.

At present, the methods of emotion recognition are mainly divided into two aspects. One is to recognize emotions through non-physiological signals such as facial expressions [5], speech [6], and body gestures [7], which can reflect the individuals' external emotional expression. Non-physiological signals are easy to collect and closely related to our daily life. Another is to identify emotions through physiological signals, such as Electroencephalogram (EEG) [8], Electromyogram (EMG) [9], Electrocardiogram (ECG) [10] and Galvanic Skin Response (GSR) [11]. Among them, the EEG signals are spontaneously electric activity of the neurons in human brain

that can reflect the truthful and plentiful emotional information within the individuals.

Because of the complementarity of different modalities during the process of emotion recognition [12]. Nowadays, many researchers begin to integrate different modalities to classify emotional states. Zheng et al. [13] proposed a multimodal emotion recognition framework called EmotionMeter, which combined the EEG signals and eye movements. Liu et al. [14] used a Deep Belief Network (DBN) to fuse speech signals and facial expressions. The emotion recognition methods based on fusing EEG signals and facial expressions have been studied in recent years. Hassouneh et al. [15] classified disabled people and Autism children's emotional states based on facial landmarks and EEG signals. Meanwhile, a classifier which combined Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models was used for classification. Zhang et al. [16] used a bimodal deep automatic encoder (BDAE) method to combine facial expressions with EEG signals. In our previous study [17], we selected the facial key points to construct texture features and extracted Power Spectral Density (PSD) features from EEG signals. The LSTM decision layer fusion strategy was utilized. Although the proposed decision level fusion method achieved better performance than single modality, but the complementarity between different modalities was ignored. Hence, in this work, we converted EEG signals into EEG topographic maps (ETM) and proposed an image-based multimodal fusion method on feature-level.

Due to the loss of a key channel during the process of emotion communication, the individuals with hearing impairment can only compensate for changes in the outside world through senses such as vision and touch. Therefore, the individuals with hearing impairment are more sensitive to emotional perception, and may have differences in recognition of emotion from healthy controls [18], [19], [20].

In this paper, in order to explore the emotional changes of individuals with hearing impairment, we designed an emotion induction experiment based on affective video clips and invited 15 participants with hearing impairment to participate in the experiment. Meanwhile, we labeled the affective video clips with discrete emotional states into the following four kinds: fear, happiness, calmness, sadness, and recorded the participates' facial expressions and EEG signals simultaneously. For signal processing, we preprocessed the collected EEG signals and facial expressions, extracted DE features of the EEG signals and converted them into the corresponding ETM. Moreover, the multichannel feature fusion method was used to fuse ETM and facial expressions, and CBAM_ResNet34 classifier was utilized to extract high-level features between different modalities and complete emotion recognition.

The layout of this article is as follows. Section II mainly introduces the related works of emotion recognition based on facial expressions and EEG signals. The whole process of the signal acquisition experiment is introduced in Section III. The feature extraction and classification model are introduced in Section IV. We carry out experimental results in Section V. We discuss this paper in Section VI. Finally, Section VII displays the conclusions.

II. Related Works

A. Emotion Recognition With Facial Expressions

Facial expressions are the most intuitive way to express human emotions. With the rapid development of deep learning in the field of image processing, researchers focus on emotion recognition based on facial expressions. Sen et al. [21] connected key facial feature points as textural features, and concatenated geometric and textural features for emotion recognition. Talele et al. [22] created a feature extraction framework called digital signature to obtain features by projecting edge pixels vertically and horizontally. Fan et al. [23] fused discriminative features extracted by CNN model with features containing shape and appearance extracted by hand. Pan et al. [24] designed a Deep Temporal–Spatial Network based on facial expressions to extract the spatiotemporal features. Chen et al. [25] proposed a facial feature called deep peak–calmness difference (DPND), which can characterize facial regions that change from calmness to expressive face, and achieved high-quality results in both unsupervised clustering and semi-supervised classification methods. In view of the superiority of deep learning methods in emotion recognition based on facial expressions, we utilize CBAM_ResNet34 to effectively extract emotion-related pixel-level features of facial expressions for emotion classification.

B. Emotion Recognition With EEG Signals

For EEG signals, researchers extracted different features and used them for emotion recognition, such as Fourier Transform (FT), Wavelet Transform (WT), PSD, Autoregressive (AR). Duan et al. [26] proposed Differential Entropy (DE) feature to characterize emotional information related to emotional states. Aydin et al. [27] proposed a new method to combine Principal Component Analysis (PCA) with Phase Space Trajectory Matrix (PSTM) for estimation of emotional features. Extracting the PCPSTM of the EEG series for short segment of 6 s. Kilic et al. [28] combined Graph Theory (GT) with short-range statistical dependency estimations for EEG analysis. Pearson Correlation (PC) was applied longer (12 sec) non-overlapped EEG segments in accordance with particular threshold as the mean. Deep learning-based algorithms have been developing in recent years, and researchers have gradually begun to extract features and complete emotion analysis based on deep learning methods. Hu et al. [29] proposed a novel convolutional layer named Scalinglayer, which can adaptively extract effective spectral features from raw EEG signals. Kawano et al. [30] used the NeuCube spiking neural network (SNN) for modeling EEG brain data related to perceiving versus mimicking facial expressions. Chao et al. [31] proposed a Capsule Network (CapsNet) model, and used the multi-feature matrix for emotion recognition. Gao et al. [32] designed a deep learning framework called channel-fused dense convolutional network, which used one-dimensional (1D) convolutional layers to extract contextual features of EEG signals. Liang et al. [33] proposed a deep convolutional recurrent generative adversarial network called EEGFuseNet to automatically extract temporal and spatial features from EEG signals. Inspired by researchers' methods on EEG signals feature extraction and classification,

we extracted representative DE features and innovatively converted DE features into ETM. CBAM_ResNet34 can capture the deeper-level features of ETM for classification.

III. Experimental Materials

In the experiment, subjects with hearing impairment were elicited with four target emotions (fear, happiness, calmness, sadness) by watching affective video clips. During the stimulation of the experimental clips, the EEG signals and facial expressions of subjects were recorded simultaneously.

A. Video Clips Selection

There are lots of manners of stimuli manner, including video clips, static pictures, and music videos. The emotion induction method based on video clips combines visual, auditory, and other sensory induction materials. The stimulation lasts a long time, which is the most effective way to induce emotions. The Queen Mary University of London used music videos to elicit emotions and proposed the DEAP [34] emotional database. Zheng et al. selected video clips as stimuli and presented emotion datasets including SEED [35], and SEED-IV [13]. However, the above databases are all based on healthy controls, not individuals with hearing impairment. In this work, we selected video clips as stimuli to induce the target emotion. The selection of happiness, sadness, and calmness clips are following the SEED database. For the selection of fearful clips, we recruited 40 postgraduates who majored in psychology. To ensure the subjects fully understood the selected video clips, each of which included subtitles. Postgraduates were asked to rate 25 fearful clips in a quiet environment, and the five highest-rated clips were finally selected. Table I presents a detailed description of the selected 20 affective video clips, each has a playback time of around 200 seconds.

TABLE I
The Description of Video Film Clips

No.	Titles of videos	Time(s)	Labels	Start time	End time
1	Lost in Thailand	238	Happiness	0:06:13	0:10:11
2	Coming Soon	192	Fear	0:01:57	0:05:09
3	World Heritage in China – 02	226	Calmness	0:00:50	0:04:36
4	Aftershock	205	Sadness	0:20:10	0:23:35
5	Back to 1942	242	Sadness	0:49:58	0:54:00
6	Coming Soon	189	Fear	1:09:13	1:12:22
7	World Heritage in China – 02	221	Calmness	0:10:40	0:13:44
8	Lost in Thailand	204	Happiness	1:05:10	1:08:29
9	Flirting Scholar	266	Happiness	1:18:57	1:23:23
10	World Heritage in China – 13	184	Calmness	0:02:59	0:06:40
11	Back to 1942	239	Sadness	2:01:21	2:05:21
12	Dead Silence	211	Fear	0:54:50	0:58:51
13	Dead Silence	182	Fear	1:15:47	1:18:49
14	World Heritage in China – 13	240	Calmness	0:10:41	0:14:41
15	Just Another Pandoras Box	241	Happiness	0:11:32	0:15:33
16	Back to 1942	240	Sadness	2:16:37	2:20:37
17	Aftershock	205	Sadness	1:48:53	1:52:18
18	the conjuring	200	Fear	1:17:42	1:21:02
19	World Heritage in China – 21	240	Calmness	0:05:36	0:09:36
20	Just Another Pandoras Box	242	Happiness	0:35:00	0:39:02

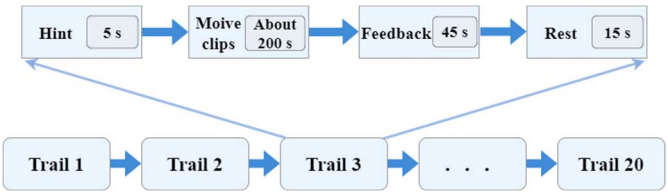

Fig. 1. The experimental stimulation procedure.

B. Subjects

Fifteen undergraduates with hearing impairment were invited to participate in this experiment. The age of the subjects was between 18 and 25, with an average age of 22, which included 12 males and 3 females. All of the subjects lost hearing in both ears, therefore, each subject was allowed to wear hearing aids during the experiment. Meanwhile, this experiment was approved by the Ethics Committee of Tianjin University of Technology.

C. Experiment Procedure

Before the experiment started, the subjects will be asked to sign an informed consent form, and make out a questionnaire for basic information, including age, specialty, cause of hearing loss, degree of hearing loss, etc. The Self-Assessment Manikin (SAM) [36] was used to rate video clips from 1 to 9, and subjects learned how to use the SAM system for self-assessment after reading the instructions and requirements of the experimental procedure.

The whole process of the experiment was shown in Fig. 1. Before formally obtaining the experiment data, the subjects were allowed to adjust to a comfortable sitting position and remain fixed. The video clips were played in sequence according to the order shown in Table I. There was a 5 seconds hint before the start of each video clip. The subject was given 45 seconds to evaluate the clip using the three dimensions of arousal, valence, and dominance. After the assessment, subjects would have a rest for 15 seconds before the next trial.

IV. Feature Extraction and Feature Classification

In this section, we preprocessed the EEG signals and facial expressions collected in the experiment. In addition, a deep learning-based classifier was proposed for multimodal fusion and emotion recognition. The framework diagram was shown in Fig. 2.

A. EEG Preprocessing and Feature Extraction

We used SymAmps2 (Neuroscan, Australia) to collect the EEG signals with 64 channels according to the international 10-20 system. The raw EEG signals were preprocessed by the EEGLAB toolbox [37]. To prevent feature redundancy, the

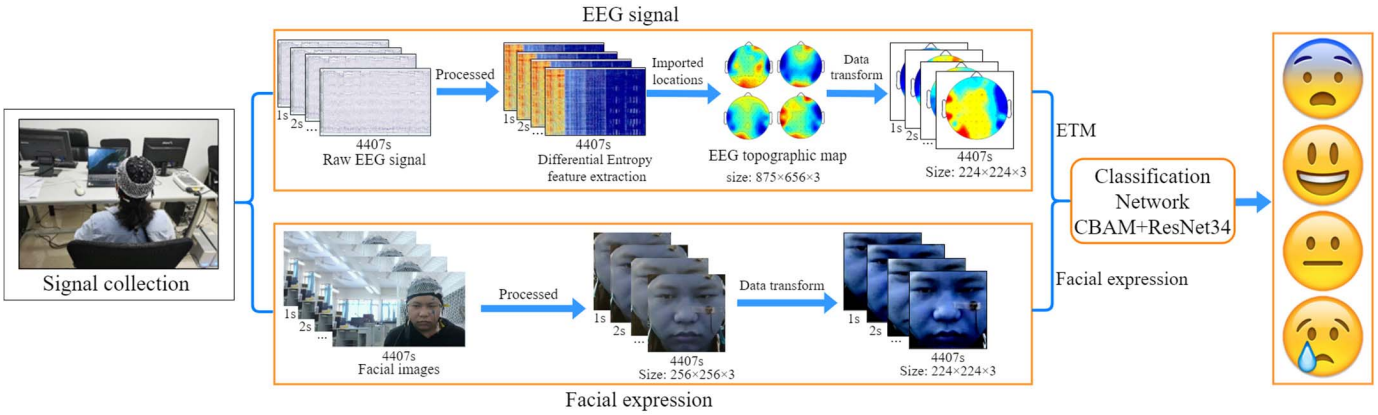

Fig. 2. The classification model framework based on combining facial expressions with ETM.

sampling frequency was reduced from 1000 Hz to 200 Hz, and we set the bilateral mastoid electrodes TP9 and TP10 as re-references to enhance the signal. The quantity of remaining electrodes is 62. Band-pass filtering from 1 to 75 Hz was used to obtain the main frequency bands related to emotions, and the power frequency interference was removed by band-pass filtering from 49 to 51 Hz. Signal artifacts were subsequently removed by Independent Component Analysis (ICA). The pre-processed EEG signals were divided into five main frequency bands (Delta: 1–3 Hz, Theta: 4–7 Hz, Alpha: 8–12 Hz, Beta: 13–30 Hz, Gamma: 31–50 Hz).

As the most commonly used frequency-domain feature, DE feature can effectively characterize emotional change. The formula for calculating the DE feature was as follows:

$$h(X) = -\int_X f(x) \log(f(x))\, dx \tag{1}$$

where X was a random variable and $f(x)$ was the probability density function of X. When the time series X obeys Gaussian distribution $N(\mu, \sigma 2)$, its DE feature can be defined as follows:

$$h(X) = \frac{1}{2} \log\left(2\pi e \sigma^2\right) \tag{2}$$

DE features were extracted by a 1-second sliding window with non-overlapping. Therefore, the total number of samples was 4407 for each subject.

ETM can use waveform and color changes to reflect the activity of different brain regions during the stimulation. The extracted DE features were converted into the corresponding ETM by the EEGLAB toolbox (image size: 875 × 656). And that was 4407 ETM for each subject. Fig. 3 shows the ETM generated by the EEGLAB toolbox, and the color map shows the intensity of brain activation, ranging from −1 to 1.

B. Facial Expression Preprocessing

During the experiment, a laptop camera (Lenovo Legion Y7000) was used to record the facial expression videos when the subjects watching the video clips at a frame rate of 25 fps. Then the videos were converted into images (image size: 1920 × 1080) by extracting the first frame per second. Furthermore, we used DLIB [38] face recognition model to remove irrelevant background images. Finally, 4407 facial expression images were obtained, which correspond with ETM on the timeline.

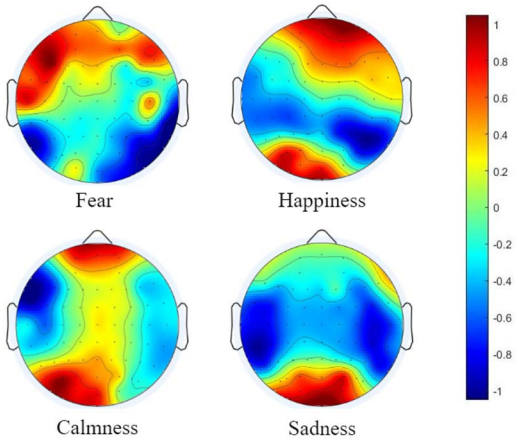

Fig. 3. EEG topographic map generated by EEGLAB toolbox.

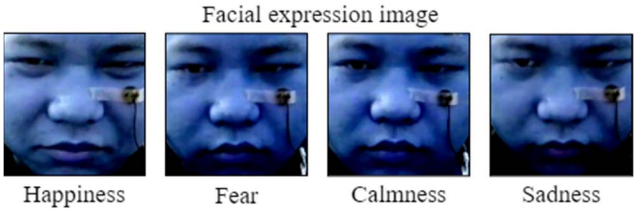

Fig. 4. Visualization of the transformed facial expression image (size: 224 × 224).

C. Classification With CBAM_ResNet34

The RGB images of ETM and facial expressions were resized to 256 × 256 pixels, and central-cropped to 224 × 224 pixels, then they were converted to a tensor format for normalization. Fig. 4 shows the transformed facial expressions.

The proposed CBAM_ResNet34 classifier was used for emotion classification. Convolutional Block Attention Module (CBAM) [39] emphasizes meaningful features through two

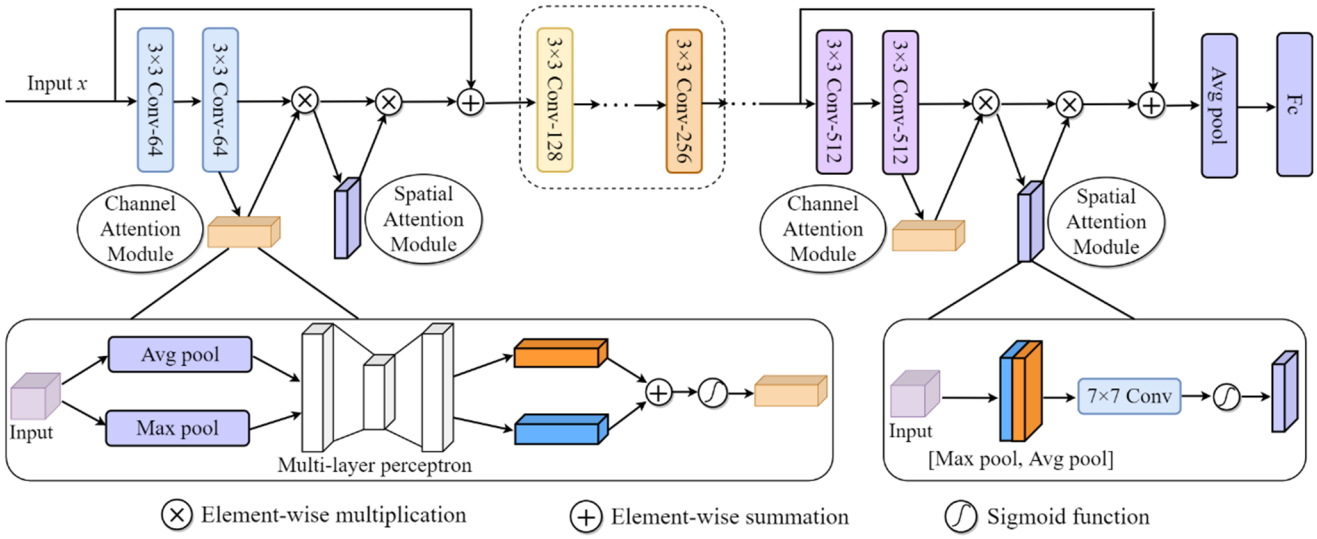

Fig. 5. The structure of the CBAM_ResNet34 classification model.

separate submodules: the channel attention module (CAM) and the spatial attention module (SAM), and sequentially obtained attention map along two independent submodules. Residual Network (ResNet) [40] was used as the backbone of the network. The residual block structure made the network not degenerate as the depth increases. The CBAM was added to every residual block of the ResNet34, the attention map was multiplied by the input feature map for adaptive refinement. The Batch Normalization (BN) layer was adopted after each convolution layer to solve the problems of gradient disappearance and gradient explosion. Fig. 5 shows the framework of the CBAM_ResNet34 model.

First, two types of feature information were obtained through the global average pooling layer and the global max pooling layer. Then, two types of feature information were fed into a multi-layer perceptron (MLP) with one hidden layer. Finally, the outputs were fused by element-wise summation, and then get channel attention value through a sigmoid activation function. We defined the channel attention value C as follows:

$$\begin{aligned} C &= \sigma\&(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{max}\right)\right)\right) \end{aligned} \tag{3}$$

where W_0 and W_1 were MLP weights, and the RELU activation function was applied on W_0. Moreover, σ represented the sigmoid function, which formula was as follows:

$$\sigma(x) = \frac{1}{1 - e^{-x}} \tag{4}$$

The features with channel attention value output from the CAM, and then the spatial attention value was extracted by the SAM. We applied average pooling and max pooling operations on the input features to highlight valid information along the channel axis. Then, the information was forwarded by a convolutional layer (kernel of size is 7×7). The spatial attention feature value named S can be expressed as follows:

$$S = \sigma\left(f\left\{\left[AvgPool(F); MaxPool(F)\right]\right\}\right) \tag{5}$$

where f represented a convolutional operation.

The most pivotal information was extracted by CBAM with CAM and SAM. Ultimately, the process of obtaining the attention feature map $F2$ through the attention module can be expressed as follows:

$$\begin{aligned} F1 &= C \otimes F \\ F2 &= S \otimes F1 \end{aligned} \tag{6}$$

where $\otimes$ represented element-wise multiplication. In the process of multiplication, the channel and spatial attention values were broadcasted. The combination of two modules allowed the neural network to know which channel and which region in the channel should be focused on during the classification process.

V. Experimental Results

A. EEG-Based Emotion Recognition

We evaluated the proposed method on a subject-dependent experiment, and the laptop Lenovo Legion Y-7000 (CPU: i5-10200H, RAM: 16G, GPU: GTX 1650Ti 4G) was used to perform programs on feature extraction and emotion classification. The experimental data of each subject was divided into training data and test data, of which the first 16 trials (3520 samples) were training data, and the last 4 trials (one for each emotion, a total of 887 samples) were test data. In the process of emotion classification based on EEG signals, 8 classifiers, which include SVM, K Nearest Neighbor (KNN), Gaussian Naive Bayesian (GNB), Random Forest (RF), Linear Discriminant Analysis (LDA), adaptive boosting (Adaboost), Logistic Regression(LR), and ResNet34, were used to compare with the proposed method. DE features of all frequency bands were used as input. For ResNet34 and CBAM_ResNet34, we put DE features of 62 channels into a 9×9 feature matrix, recognized as the input of the classifier. 64 was the number of the batch size, and the AdamW was used as the optimizer to speed up the convergence of the classifier. Besides, we used the Cross-Entropy as a loss function. RELU was selected

TABLE II
THE PARAMETER SELECTION FOR DIFFERENT CLASSIFIERS

Classifiers	Parameters
SVM	Linear kernel
	C = 0.5
KNN	K = 2
NB	Gaussian NB
RF	N estimators = 7
LDA	Solver = svd
	N components = 2
Adaboost	Decision Tree Classifier
	Learning rate: 0.1
	N estimators = 100
LR	Solver = liblinear
	Penalty = l2
ResNet34 CBAM_ResNet34	Learning rate: 0.001 Scheduler: CosineAnnealingLR Optimizer: AdamW Activation function: RELU Loss function: Cross entropy

TABLE III
PERFORMANCE ANALYSIS OF DE FEATURES USING STATISTICAL PARAMETERS WITH SVM, KNN, GNB, RF, LDA, LR, RESNET34 AND CBAM_RESNET34

Classifier Types	All band			
	Accuracy	Precision	F1-score	Sensitivity
SVM	44.32	47.04	41.62	44.13
KNN	34.63	37.02	33.29	34.64
GNB	36.36	35.55	33.40	35.80
RF	35.26	35.60	33.61	35.11
LDA	48.28	51.78	43.50	47.81
Adaboost	35.06	38.14	32.03	35.03
LR	48.75	52.62	45.06	48.38
ResNet34	53.32	54.21	53.60	53.01
CBAM_ResNet34	**58.86**	**59.77**	**59.17**	**58.59**

as the activation function to further avoid the problems of gradient saturation and gradient disappearance. The number of the epoch is 50. To avoid getting stuck in a local optimum during the verification process, we used cosine annealing to control the learning rate and set the original learning rate to 0.001 as well as decreasing once every iteration. Table II showed the parameters of different classifiers.

Table III showed the performance of different classifiers. It can be seen that the classification accuracy, F1-score, Precision and Sensitivity metrics of CBAM_ResNet34 classifier have reached better results, which were significantly improved compared with traditional machine learning classifiers. The proposed CBAM_ResNet34 classifier get an outstanding performance of 58.86% for four emotions classification. At the same time, it also demonstrated that the deep learning method can better capture the emotion-related feature, and the recognition was more efficient and accurate.

To further explore the classification accuracy of different classifiers on the all-frequency band, Fig. 6 shows the result of fifteen subjects using all 9 classifiers. Compared with other classifiers, the CBAM_ResNet34 classifier always maintains the best recognition accuracy, and the accuracy of subject 5 can achieve more than 80%. Meanwhile, it can be seen from the results that there are large differences in the classification

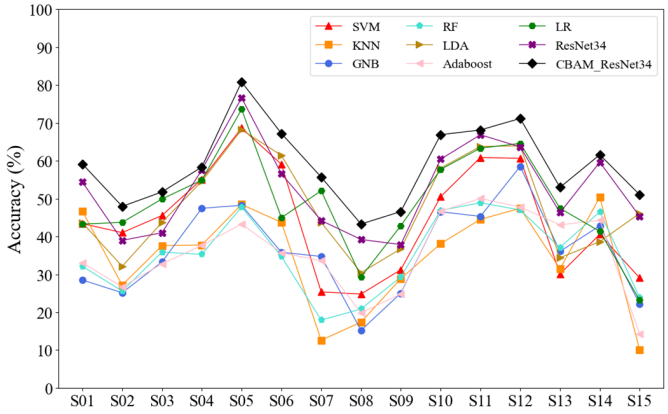

Fig. 6. The different classifier performance of fifteen subjects with hearing impairment on all-frequency band.

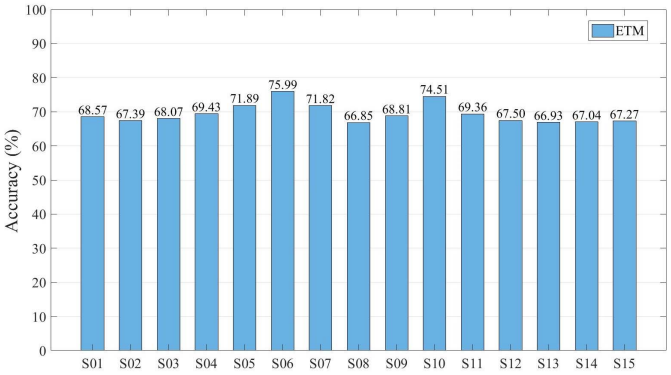

Fig. 7. Classification accuracy on ETM of fifteen subjects using the CBAM_ResNet34 classifier.

results of different subjects using the DE features for classification whether it is a machine learning or deep learning classifier.

After emotion analysis on all band DE features, we converted the DE features into corresponding ETM and used CBAM_ResNet34 to learn EEG topographical representations. Fig. 7 showed the classification accuracy on the ETM of fifteen subjects. The highest recognition accuracy of subject 6 is 75.99%, and the lowest recognition accuracy of subject 8 is 66.85%. Compared with using DE features for emotion classification, ETM not only has better recognition performance but also has less variability among all subjects.

B. Facial Expression Emotion Recognition

The deep learning method integrates feature extraction and emotion classification, using convolution kernels of different sizes to capture emotion-related feature, optimizing the error through backpropagation to improve the accuracy. In this work, the raw facial expressions were used as input to the classifier. Fig. 8 shows the classification results of fifteen subjects based on facial expressions, the highest performance was subject 6 (71.48%), and the lowest performance was subject 2 (63.61%). Fig. 8 took subject 7 as an example to show the feature map of different layers output, (a), (b), (c), and (d) represented of layer1 to layer4. From Fig. 9 (b), it can

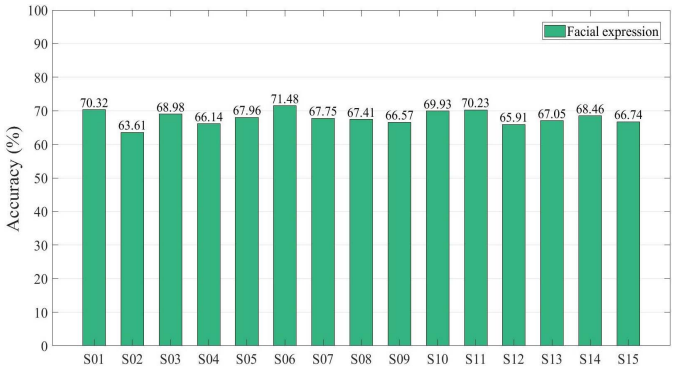

Fig. 8. Classification accuracy on facial expressions of fifteen subjects using the CBAM_ResNet34 classifier.

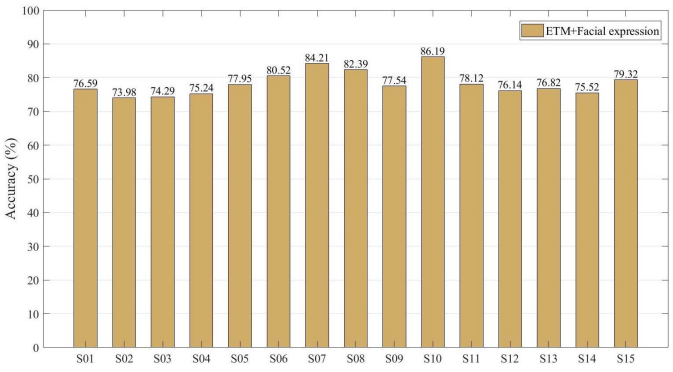

Fig. 10. Classification accuracy of fifteen subjects fused by ETM and facial expressions.

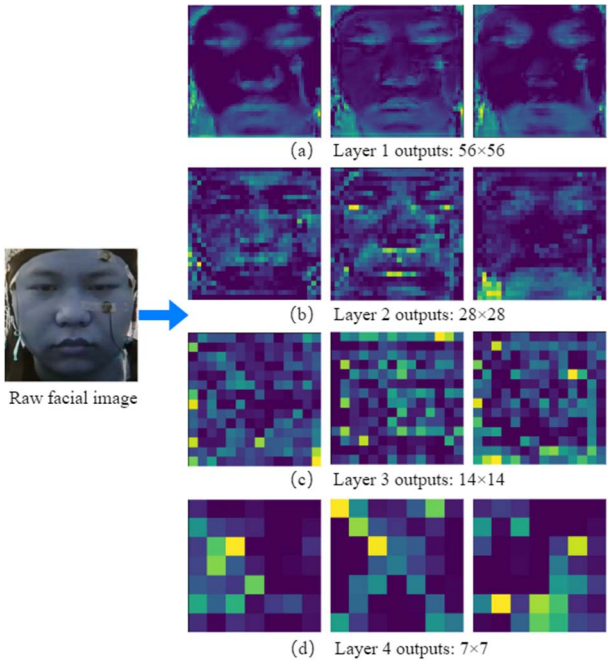

Fig. 9. The feature map of different layers output of subject 7, (a) layer 1 outputs, (b) layer 2 outputs, (c) layer 3 outputs, (d) layer 4 outputs.

be seen that the emotion-related regions concerned by the classification network were more extensive and comprehensive than the traditional facial landmark localization, and it can further lock the regions related to emotions through deeper convolutional layers.

C. Fusion of ETM and Facial Expressions

We converted the DE features into ETM and fused the facial expressions of the corresponding segment. Finally, the ETM and the facial expressions, both 2-dimensional (2D) feature matrices, were fused and fed into the CBAM_ResNet34 classifier for emotion recognition. Fig. 10 showed the recognition performance based on multimodal fusion. In the case of emotion recognition using fusion features, the average accuracy on the database of subjects with hearing impairment was achieved at 78.32%. Compared with facial expression, the accuracy increased by 10.42%, and compared with ETM, the accuracy increased by 8.89%. The classification accuracy of each subject after multimodal fusion was higher than that of any single modality. These results showed that combined facial expressions and ETM can significantly improve the recognition performance.

TABLE IV
COMPARISON OF THE MEAN ACCURACY BASED ON THREE SINGLE MODALITY AND TWO MULTIMODAL

Categorical Features		Accuracy (%)
Single Modality	Facial expression	67.90
	DE	58.86
	ETM	69.43
Multimodal	DE + Facial	68.49
	ETM + Facial	78.32

VI. DISCUSSION

To obtain emotional representational features between different modalities, we proposed a feature-level fusion strategy that combined ETM and facial expressions. Table IV shows the average recognition accuracy based on facial expressions, DE features, ETM, fused DE features with facial expressions, and fused ETM with facial expressions. The results showed that the proposed method of using ETM for emotion recognition was better than DE features, and the performance of combining facial expressions with ETM was higher than that of combining facial expressions with DE features.

To further explore the complementary characteristics of facial expressions and ETM during the process of classification, we analyzed the confusion matrix to explore the performance of different modalities in recognizing different emotions. Here we listed the confusion matrix of subject 7 as an example. Fig. 11 (a, b, c) showed the confusion matrix of subject 7's facial expressions, ETM, and multimodal fusion. The confusion matrix showed that ETM had the best performance in recognizing fear emotions, and facial expressions had more advantages in recognizing happiness and sadness emotions. In the recognition process based on two single modality, there were many cases of misclassification. However, after fusing facial expressions and ETM, the misclassification was reduced and the robustness of recognition was improved. Meanwhile, after multimodal fusion, the recognition performance of calmness was significantly improved. To further

TABLE V
COMPARISON OF THE PROPOSED SCHEME WITH STATE OF THE ART METHODS FOR HUMAN EMOTION RECOGNITION

Authors	Labels	Database	Feature extraction method	Classifier	Accuracy (%)
Chen et al. [41]	Positive, neutral, negative	SEED	DE	LR	77.40
Chao et al. [42]	Arousal, Valence, Dominance	DEAP	Multiband feature matrix	Capsule Network	Arousal: 68.28 Valence: 66.73 Dominance: 67.25
Arjun et al. [43]	Positive, neutral, negative	SEED	LSTM with Channel Attention Autoencoder	CNN with Attention	76.70
Yang Li et al. [44]	Sad, fear, happy, neutral	SEED IV	Raw EEG signals	Bi-hemispheric discrepancy model	69.03
Zhong et al. [45]	Sad, fear, happy, neutral	SEED IV	Raw EEG signals	Graph Neural Network	73.84
Zhu et al. [46]	Sad, fear, happy, neutral	SEED IV	Autoencoder	Multisource Wasserstein Adaptation Coding Network	76.04
Our method	Fear, happiness, calmness, sadness	Own database	ETM + Facial expression	CBAM_ResNet34	**78.32**

Fig. 11. Confusion matrices of subject 7 including two single modality and multimodal fusion. Each row of the confusion matrix represents the true kind of the sample, and each column represents the predicted kind. (a) ETM, (b) Facial expressions, (c) Multimodal fusion.

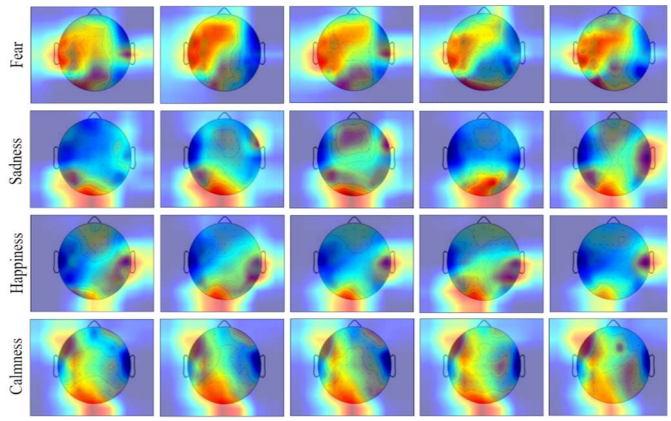

Fig. 12. Five different time slices ETM of subject 7 were randomly selected in chronological order to construct the Grad-CAM.

explore the key regions of the brain that affect the emotion of subjects with hearing impairment, we constructed the Gradient-weighted Class Activation Mapping (Grad-CAM) [47] for analyzing. Fig. 12 showed subject 7's Grad-CAM constructed by five-time slices of four target emotions randomly selected in chronological order. The brain regions influenced by emotion were primarily located in the prefrontal, temporal, and occipital lobes.

We have compared our proposed method with state-of-the-art techniques, Table V summarizes the main studies that classify human emotions. The proposed method that combined ETM and facial expressions was computationally efficient and also achieves an outstanding accuracy of 78.32% for four emotions classification.

VII. CONCLUSION

In this paper, a multimodal emotion recognition framework was proposed to classify four kinds of emotion (fear, happiness, calmness, sadness) of subjects with hearing impairment. In order to solve the disadvantage of EEG signals' nonlinearity and nonstationary characteristics during the multimodal fusion process. For EEG signals, DE features were converted into ETM. We fused the ETM and facial expressions, and the CBAM_ResNet34 was present to extract and classify emotional representational information. The subject-dependent experimental results showed that the classification accuracy of the proposed strategy achieved the 78.32%, which was better than the single modality (DE, 58.86%, ETM, 69.43%, and facial expressions, 67.90%) and combined DE with facial expressions (68.49%). This may indicate that the fusion of ETM with facial expressions can effectively explore the complementarity between different modalities during the process of emotion recognition. We also demonstrated the validity of the proposed method by comparing the method with state-of-the-art works. In the future, we will optimize the classification algorithm and preprocessing process to develop a cross-subject classification model.

REFERENCES

[1] N. Guo et al., "SSVEP-based brain computer interface controlled soft robotic glove for post-stroke hand function rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1737–1744, 2022.
[2] C. Jiang, Y. Li, Y. Tang, and C. Guan, "Enhancing EEG-based classification of depression patients using spatial information," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 566–575, 2021.
[3] P. Ekman and W. V. Friesen, "Facial action coding system (FACS): A technique for the measurement of facial actions," *Rivista Di Psichiatria*, vol. 47, no. 2, pp. 38–126, 1978.
[4] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[5] M. Mukhopadhyay, A. Dey, R. N. Shaw, and A. Ghosh, "Facial emotion recognition based on textural pattern and convolutional neural network," in *Proc. IEEE 4th Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Sep. 2021, pp. 6–11.
[6] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.
[7] F. Ahmed, A. S. M. H. Bari, and M. L. Gavrilova, "Emotion recognition from body movement," *IEEE Access*, vol. 8, pp. 11761–11781, 2020.
[8] M. A. Ozdemir, M. Degirmenci, E. Izci, and A. Akan, "EEG-based emotion recognition with deep convolutional neural networks," *Biomed. Eng./Biomedizinische Technik*, vol. 66, no. 1, pp. 43–57, Feb. 2021.
[9] Y. Cheng, G.-Y. Liu, and H.-L. Zhang, "The research of EMG signal in emotion recognition based on TS and SBS algorithm," in *Proc. 3rd Int. Conf. Inf. Sci. Interact. Sci.*, Jun. 2010, pp. 363–366.
[10] T. Dissanayake, Y. Rajapaksha, R. Ragel, and I. Nawinne, "An ensemble learning approach for electrocardiogram sensor based human emotion recognition," *Sensors*, vol. 19, no. 20, pp. 1–24, 2019.
[11] M. Liu, D. Fan, X. Zhang, and X. Gong, "Human emotion recognition based on galvanic skin response signal feature selection and SVM," in *Proc. Int. Conf. Smart City Syst. Eng. (ICSCSE)*, Nov. 2016, pp. 157–160.
[12] Z. He et al., "Advances in multimodal emotion recognition based on brain-computer interfaces," *Brain Sci.*, vol. 10, no. 10, pp. 1–29, 2020.
[13] 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.
[14] D. Liu, L. Chen, Z. Wang, and G. Diao, "Speech expression multimodal emotion recognition based on deep belief network," *J. Grid Comput.*, vol. 19, no. 2, pp. 1–13, Jun. 2021.
[15] 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," *Informat. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100372.
[16] H. Zhang, "Expression-EEG based collaborative multimodal emotion recognition using deep AutoEncoder," *IEEE Access*, vol. 8, pp. 164130–164143, 2020.
[17] D. Li et al., "The fusion of electroencephalography and facial expression for continuous emotion recognition," *IEEE Access*, vol. 7, pp. 155724–155736, 2019.
[18] Y. Yang, Q. Gao, Y. Song, X. Song, Z. Mao, and J. Liu, "Investigating of deaf emotion cognition pattern by EEG and facial expression combination," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 2, pp. 589–599, Feb. 2022.
[19] Q. Kang et al., "Emotion recognition from EEG signals of hearing-impaired people using stacking ensemble learning framework based on a novel brain network," *IEEE Sensors J.*, vol. 21, no. 20, pp. 23245–23255, Oct. 2021.
[20] Z. Tian, D. Li, Y. Song, Q. Gao, Q. Kang, and Y. Yang, "EEG-based emotion recognition of deaf subjects by integrated genetic firefly algorithm," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.
[21] D. Sen, S. Datta, and R. Balasubramanian, "Facial emotion classification using concatenated geometric and textural features," *Multimedia Tools Appl.*, vol. 78, no. 8, pp. 10287–10323, Apr. 2019.
[22] K. Talele and K. Tuckley, "Facial expression recognition using digital signature feature descriptor," *Signal, Image Video Process.*, vol. 14, no. 4, pp. 701–709, Jun. 2020.
[23] X. Fan and T. Tjahjadi, "Fusing dynamic deep learned features and handcrafted features for facial expression recognition," *J. Vis. Commun. Image Represent.*, vol. 65, Dec. 2019, Art. no. 102659.
[24] X. Pan et al., "Video-based facial expression recognition using deep temporal–spatial networks," *IETE Tech. Rev.*, vol. 37, no. 4, pp. 402–409, Jul. 2020.
[25] J. Chen, R. Xu, and L. Liu, "Deep peak-neutral difference feature for facial expression recognition," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29871–29887, Nov. 2018.
[26] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Nov. 2013, pp. 81–84.
[27] S. Aydin, "Deep learning classification of neuro-emotional phase domain complexity levels induced by affective video film clips," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1695–1702, Jun. 2020.
[28] B. Kılıç and S. Aydın, "Classification of contrasting discrete emotional states indicated by EEG based graph theoretical network measures," *Neuroinformatics*, vol. 20, no. 4, pp. 863–877, Oct. 2022.
[29] J. Hu, C. Wang, Q. Jia, Q. Bu, R. Sutcliffe, and J. Feng, "ScalingNet: Extracting features from raw EEG data for emotion recognition," *Neurocomputing*, vol. 463, pp. 177–184, Nov. 2021.
[30] H. Kawano, A. Seo, Z. G. Doborjeh, N. Kasabov, and M. G. Doborjeh, "Analysis of similarity and differences in brain activities between perception and production of facial expressions using EEG data and the NeuCube spiking neural network architecture," in *Proc. Int. Conf. Neural Inf. Process.*, vol. 9950, 2016, pp. 221–227.
[31] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion recognition from multiband EEG signals using CapsNet," *Sensors*, vol. 19, no. 9, p. 2212, May 2019.
[32] Z. Gao, X. Wang, Y. Yang, Y. Li, K. Ma, and G. Chen, "A channel-fused dense convolutional network for EEG-based emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 4, pp. 945–954, Dec. 2021.
[33] Z. Liang et al., "EEGFuseNet: Hybrid unsupervised deep feature characterization and fusion for high-dimensional EEG with an application to emotion recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1913–1925, 2021.
[34] S. Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.
[35] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.
[36] J. D. Morris, "OBSERVATIONS: SAM: The self-assessment manikin—An efficient cross-cultural measurement of emotional response," *J. Advert. Res.*, vol. 35, no. 6, pp. 63–68, 1995.
[37] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.
[38] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jul. 2009.
[39] S. Woo et al., "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 770–778.
[41] D.-W. Chen et al., "A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition," *Sensors*, vol. 19, no. 7, p. 1631, Apr. 2019.
[42] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion recognition from multiband EEG signals using CapsNet," *Sensors*, vol. 19, no. 9, p. 2212, May 2019.
[43] Arjun, A. S. Rajpoot, and M. R. Panicker, "Subject independent emotion recognition using EEG signals employing attention driven neural networks," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103547.
[44] Y. Li et al., "A novel bi-hemispheric discrepancy model for EGG emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 2, pp. 354–367, Jun. 2021.
[45] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul. 2022.
[46] L. Zhu et al., "Multisource Wasserstein adaptation coding network for EEG emotion recognition," *Biomed. Signal Process. Control*, vol. 76, Jul. 2022, Art. no. 103687.
[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019.