Received: 25 July 2023 | Revised: 16 November 2023 | Accepted: 13 December 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12346

ORIGINAL RESEARCH

Residual multimodal Transformer for expression-EEG fusion continuous emotion recognition

Xiaofang Jin[1] | Jieyu Xiao[1] | Libiao Jin[1] | Xinruo Zhang[2]

[1]College of Information and Communication Engineering, Communication University of China, Beijing, China

[2]School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

Correspondence

Jieyu Xiao, College of Information and Communication Engineering, Communication University of China, Beijing 100024, China.
Email: xiaojy177@163.com

Funding information

The State Key Development Program in 14th Five-Year, Grant/Award Number: 2021YFF0900701

Abstract

Continuous emotion recognition is to predict emotion states through affective information and more focus on the continuous variation of emotion. Fusion of electroencephalography (EEG) and facial expressions videos has been used in this field, while there are with some limitations in current researches, such as hand-engineered features, simple approaches to integration. Hence, a new continuous emotion recognition model is proposed based on the fusion of EEG and facial expressions videos named residual multimodal Transformer (RMMT). Firstly, the Resnet50 and temporal convolutional network (TCN) are utilised to extract spatiotemporal features from videos, and the TCN is also applied to process the computed EEG frequency power to acquire spatiotemporal features of EEG. Then, a multimodal Transformer is used to fuse the spatiotemporal features from the two modalities. Furthermore, a residual connection is introduced to fuse shallow features with deep features which is verified to be effective for continuous emotion recognition through experiments. Inspired by knowledge distillation, the authors incorporate feature-level loss into the loss function to further enhance the network performance. Experimental results show that the RMMT reaches a superior performance over other methods for the MAHNOB-HCI dataset. Ablation studies on the residual connection and loss function in the RMMT demonstrate that both of them is functional.

KEYWORDS

facial expression recognition, human-machine interaction, information fusion, physiology, regression analysis

1 | INTRODUCTION

Continuous emotion recognition refers to the continuous identification of emotional categories or prediction of valence-arousal values over time. The discrete categories of emotion identification initially came from the six basic emotions theory proposed by Ekman [1] in 1992. However, human emotions are always complex and diverse, so many researchers choose the dimensional model proposed by Lang [2]. In the dimensional model, each emotion can correspond to a set of valence-arousal values, so more emotional categories can be covered, which makes it more convenient to distinguish similar emotions. This paper takes the dimensional model as the target of emotion analysis and continuously analyses sentiment over time.

Facial expressions, as the most intuitive and observable form of emotional expression, have been extensively employed in sentiment analysis. Initially, the extraction of facial action units or landmark features [3, 4] is required in facial expression recognition. However, with end-to-end models such as convolutional neural networks (CNN), the complex process of manual feature extraction can be skipped, and emotions can be recognised directly through inputs. 2D CNN plays an important role in image sentiment recognition, while it lacks expression continuity information. 3D CNN can be used for recognising emotions in videos, but the model becomes difficult to be trained due to its large number of parameters. Whereas using a combination of 2D CNN and 1D CNN as a substitute for 3D CNN is feasible, which have demonstrated by studies such as in refs. [5, 6]. In this paper, we also utilise a

2D spatial feature extractor and a 1D temporal feature extractor to extract the spatiotemporal information of the videos.

Facial expressions are subjectively controlled and can be easily disguised. If the subjects are unable to accurately express their emotional states or subjectively conceal their emotions, the accuracy of sentiment analysis will be affected. In this case, electroencephalography (EEG) and other physiological signals that are less influenced by subjective factors can compensate for the insufficient information source, and are more conducive to sentiment analysis. The brain is a high-level neural centre that controls movement, generates sensation and enables advanced cognitive functions. Du [7] decoded visual neural representations from EEG, further demonstrating the abundance of EEG information. Hence, EEG has been widely employed in sentiment analysis with three types of features including time-domain features [8], frequency-domain features [9], and time-frequency features [10]. In this paper, EEG is represented with the most commonly used feature type, frequency-domain power spectral density (PSD).

As human emotions are inherently multimodal, one of the key research focuses is how to effectively integrate multiple emotional features, such as facial expressions, voice, physiological signals, and others. Some multimodal fusion methods do not rely on specific models, such as feature fusion, decision fusion etc., while others integrate the process of fusing different modalities directly into the deep learning network [11]. The former is widely adopted, for example, Yin [12] fused music and skin conductance using feature concatenation, achieving an accuracy of 83.76%, which is a 7% improvement compared to the single-modality approach. The latter includes Wu [13] who utilised long short-term memory (LSTM) and temporal self-attention mechanism to fuse EEG and facial expression signals for emotion recognition, achieving at least 2% higher accuracy than other models. Similarly, Tsai [14] proposed a multimodal Transformer that integrates text, audio, and video modalities, and improves the accuracy of emotion recognition on CMU-MOSEI dataset by 10%–15%. Due to the significant advantages of Transformer in various fields of deep learning, we propose a framework named residual multimodal Transformer (RMMT) using the multimodal Transformer based on the multimodal attention mechanism to fuse facial expressions and EEG. In addition, a residual connection module is added to fuse shallow and deep features, further improving the accuracy of continuous emotion recognition. Furthermore, inspired by knowledge distillation [15–17], we add a L1 loss between the spatiotemporal features of EEG and facial expressions in the loss function to fully exploit the advantages of the two modalities and train a model that is more favourable for emotion analysis.

Facial expression is the most important external feature of emotional expression, and EEG is the most related physiological signal to emotions. Through the complementarity and fusion of these two modes in the RMMT, the accuracy of continuous emotion recognition is improved. The experiments are conducted on the MANHOB-HCI dataset, and the results demonstrate the effectiveness of the proposed approach in enhancing emotion analysis.

This paper makes two main contributions:

1) A new continuous emotion recognition model based on the fusion of EEG and facial expressions is constructed, utilising cross-modal attention mechanisms for multimodal fusion. Moreover, informed by the principles of knowledge distillation, loss between spatiotemporal features of the two modalities, namely KD loss is added into the loss function. Extensive experiments and comparisons are conducted on MANHOB-HCI dataset, and the results demonstrate that the proposed method consistently outperforms the state-of-the-art methods by a large margin.
2) A residual connection is added on the basis of the multimodal Transformer to combine shallow and deep features, which further improved the effectiveness of affective analysis.

2 | RELATED WORKS

2.1 | Multimodal emotion analysis

In interpersonal communication, humans often use various modalities such as languages and facial expressions to accurately convey information and express emotions. D'mello [18] indicates that multimodal systems consistently outperform single modality systems, with an average accuracy improvement of 9.83%. Du [19] solved the labelled-data-scarcity problem and the missing-modality problem at the same time through a novel multi-view deep generative framework which is a very successful method to fuse multiple modalities.

According to the different stages of fusion, multimodal fusion can be classified into feature fusion, decision fusion, and hybrid fusion. Their process is presented in Figure 1. Feature fusion, also known as early fusion, refers to the fusion at the data or feature level. It only requires the training of a single model. While one or more modalities are missing, such early fusion will fail. Decision fusion is typically performed after the single-modal output results are obtained. It is more flexible and superior because the optimal classifiers can be selected for different modalities and it can still work with some missing modalities. However, the inter-modality correlations remain unutilised. Hybrid fusion is a method that combines feature and decision fusion approaches and exploits the advantages of them while it requires more parameters and computational resources. In 2022, Sun [20] fused brain functional connectivity networks and eye gaze at feature level and achieved a classification accuracy of 91.32%. Tian [21] achieved the optimal fusion of facial expressions and audio signals through Bayesian that also known as decision fusion method. The accuracy of the model reached 98.56% on the CK + dataset. A hybrid fusion approach was used by Ayari [22] to integrate the prediction emotions from text, audio, and facial expressions with features from event detection model. The performance of

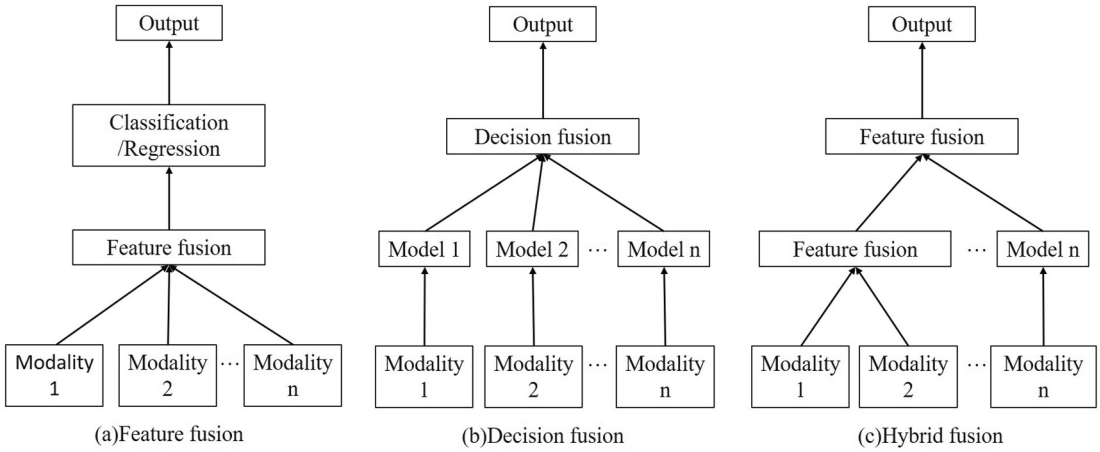


FIGURE 1 Diagram of different fusion methods.

multimodal emotion recognition based on hybrid fusion was enhanced in terms of F1 score from 0.67 to 0.95.

The fusion methods mentioned above are agnostic to the model architecture and the input to the model remains a feature vector, which is either a unimodal feature or a new concatenated feature vector composed of multiple modalities. The fusion methods based on models are distinct from the above fusion methods, because they are specifically proposed to handle multi-modal information [11]. In these approaches, the modalities are no longer fused by simple methods like feature concatenation before inputted models, instead, their features are directly inputted models. Deep neural networks approaches are the typical representatives and show good performance when compared to non-neural networks with their capacity to learn from large amounts of data. However, a common challenge encountered when training deep learning networks is the lack of sufficient data. In 2021, Zhou [23] proposed a multimodal fusion attention network based on adaptive multilevel factorised bilinear pooling for audio-visual emotion recognition, which achieved an accuracy of 75.49% on the IEMOCAP dataset outperforming previous models. In 2022, Wang [24] utilised CNN to extract spatial features of EEG and used the statistical features, approximate entropy, and hurst exponents as temporal features. Then they fused these features through Bi-LSTM to recognise emotion, and finally achieved outstanding performance. In this paper, a model which realises the fusion of EEG and facial expressions based on the RMMT is proposed, fully leveraging the advantages of feature fusion.

2.2 | Continuous emotion recognition

Discrete emotion recognition is limited in expressing a finite set of emotional categories and neglects the temporal continuity of emotional changes, while continuous emotion recognition more focuses on the changes. In spatial terms, emotional states are treated as continuous point values in a multidimensional emotional space, while in temporal terms, emotional states and their changing trends are predicted continuously over a certain time interval ultimately. Continuous emotion recognition presents more obstacles and challenges to overcome compared to discrete emotion recognition.

In 2020, Pei [25] demonstrated the effectiveness of 3D morphable models in extracting spatiotemporal information of facial muscle deformations for continuous emotion recognition. It outperformed handcrafted features by average of 37.05% on the RECOLA development set. Chen [26] employed a CNN and multi-head attention to capture dynamic relationships both between and within audio and video modalities and reached 0.583 for arousal and 0.564 for valence in terms of Consistency Correlation Coefficient (CCC) on the AVEC2019 dataset, achieving state-of-the-art performance. In 2021, Hu [27] utilised a combination of 2D CNN and 1D CNN for continuous video emotion analysis, with static features of image sequences extracted by a residual attention network and then input into a two-stage spatiotemporal attention time convolutional network. The model achieved CCC values of 0.659 and 0.69 for arousal and valence on the RECOLA dataset, respectively. These results represent an improvement of 25%–36% compared to Pei's results. In 2022, Li [28] proposed a multi-timescale model and verified that combining time pooling at different scales can fully utilise temporal information and improve the accuracy of continuous emotion recognition.

The majority of existing literatures on continuous emotion recognition primarily concentrate on audio and video modalities, with relatively fewer based on the fusion of EEG and facial expressions. In 2016, Soleymani [29] proposed the MANHOB-HCI dataset and extracted PSD features from EEG and landmark features from facial expressions. They used and compared various basic models, including multilayer perceptron, support vector machine, conditional random field, and long-short term memory. But ref. [29] only simply concatenated the features. Similarly, Li [30] also used PSD features from EEG signals and landmark features from facial expressions in 2019. Furthermore, in EEG signal processing, t-SNE was utilised for feature selection (dimension reduction) to

simplify the model and enhance its generalisation ability. Then, a multi-step LSTM was adopted to fuse the two modalities, and ultimately, the model attained desirable outcomes on self-collected data. In 2020, Choi [31] proposed a fusion network that combined CNN and LSTM to extract spatiotemporal features from both EEG and facial expressions modalities. Multiple layers of attention structures based on bilinear pooling were used to fuse the two modalities, and the proposed fusion network improved the performance by at least 6.9% compared to single-modal models. But bilinear pooling method may ignore the correlation between different modalities, which may lead to the loss of critical information. In 2022, Zhang [32] utilised knowledge distillation to continuous emotion recognition based on EEG and facial expressions modalities. They trained the EEG feature extraction model with the more effective facial expressions modality to improve the information utilisation rate of the EEG modality. The results indicate that with the assistance of the facial expressions, the EEG feature extraction model can extract more favourable features for emotion analysis, thereby improving the accuracy of the EEG unimodal emotion analysis. But facial expressions are not involved in the process of emotion prediction.

In the field of continuous emotion recognition combining EEG and facial expressions, the PSD, a typical feature of EEG, has never been abandoned, while as for facial expressions, the commonly used features have evolved from handcrafted features to deep learning features. Hence, while retaining the PSD features, we employ a CNN to extract facial expressions features. Building upon the advantages of knowledge distillation and multimodal fusion, the loss function is optimizsed and a RMMT is employed in fusing EEG and facial expressions to achieve the best performance.

3 | PROPOSED METHOD

The proposed continuous emotion recognition model RMMT is illustrated in Figure 2. There are four components in the model: a face feature extractor, an EEG feature extractor, a feature fusion module and a regressor. In the face feature extractor, image sequences from the videos are fed into a pre-trained 50-layer residual network (ResNet50) [33] and a temporal convolutional network (TCN) [34] to obtain spatiotemporal features. In the EEG feature extractor, after preprocessing such as denoising and re-reference, EEG signals are first processed in the PSD calculation to calculate the relative power density of each frequency band, and then fed into a TCN to extract the spatiotemporal features of EEG. Subsequently, the spatiotemporal signals of the two modalities are fused by a feature fusion module based on a multimodal Transformer with a residual connection module. Finally, the emotional states are predicted by a regressor.

The details of each module will be introduced in the following subsections.

3.1 | Face feature extractor

As shown in Figure 2, to extract spatial and temporal features of videos respectively, a pre-trained ResNet50 and a TCN are utilised in the face feature extractor. The ResNet50 network is first pre-trained on the MS-CELEB-1M dataset for facial recognition as a downstream task, and subsequently fine-tuned on the FER + dataset for facial expressions analysis. In modelling sequential data, the most commonly used method is recurrent neural network (RNN) and its variants, such as LSTM and gated recurrent unit. However, TCN is a completely different model for analysing time-series data based on dilated causal convolutions. Causal convolution, similar to LSTM is a unidirectional model of which output of each moment only relates to the previous layers and values while the future data is unknown for it. Such a model can only capture a fixed number of values over time. To capture information over longer time periods, more network layers need to be added, which not only increases the number of parameters but also makes the model harder to train. Thus, dilated convolution is proposed. Dilated convolution allows for interval subsampling of the input: a subsampling distance of 1 means that every point of the input is sampled, while a subsampling distance of two means that every second point is sampled. The greater the depth of the network, the larger the available sampling distance, thus, the effective window size grows exponentially with the depth,

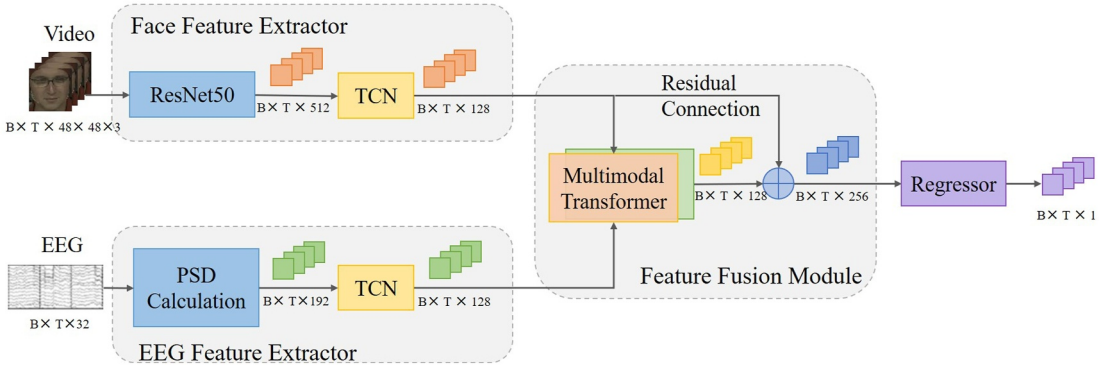


FIGURE 2 The illustration of the RMMT. EEG, electroencephalography; PSD, power spectral density; RMMT, residual multimodal Transformer; TCN, temporal convolutional network.

thereby obtaining a larger receptive field with fewer network layers. According to the research, TCN composed of dilated causal convolutions has achieved better accuracy than RNN in a variety of fields, hence, in this paper, TCN is utilised as the main network for extracting temporal features.

3.2 | EEG feature extractor

Similar to the face feature extractor, the EEG feature extractor is composed of a spatial feature extractor and a temporal feature extractor as illustrated in Figure 2. The latter is a TCN same to the facial feature extraction module, while the former is a relative PSD calculation which has been widely used in EEG emotion recognition. The computation of relative power spectrum for EEG data serves as a normalisation step that enhances the accuracy.

EEG signals can typically be divided into six frequency bands, with each band associated with a distinct mental state, which is summarised in Table 1. The δ wave (0.3–5 Hz) is characterised by high-amplitude waves and is associated with the unconscious mind. It often appears during sleeping or unconscious states and is typically located in the frontal region in adults and the occipital region in infants. The θ wave (5–8 Hz) is associated with dreaming and is also observed during meditation. The α wave (8–12 Hz) is associated with a relaxed but conscious mental state and is the most notable rhythmical brain wave that can be detected in both the posterior and lateral regions of the brain. The β wave (12–30 Hz) is a low-amplitude wave that often fluctuates with brain activity. It typically occurs symmetrically on both sides of the brain and is most prominent in the frontal region. The γ wave (30–45 Hz) generally appears in the somatosensory cortex, and its presence is usually associated with highly excited or stimulated states. The β wave can be further divided into β_1 (12–18 Hz) and β_2 (18–30 Hz), with the former associated with a relaxed state of focused attention, while the latter indicates a state of deep concentration and high mental effort.

Based on the theoretical knowledge above, the relative PSD of 32 channels of EEG in 6 frequency bands is calculated and the feature dimension is $L \times 192$ (L represents the number of sampling points, $192 = 32 \times 6$). The periodogram is the simplest spectral estimation technique for stochastic signals like EEG. The method is a biased estimation, but results in an uneven power spectrum with a large mean square error and hence a non-consistent estimate. Due to the limitations of the periodogram, the Welch method is employed to estimate the PSD. The data $x(n)$ with a length of N ($n = 0, 1, \ldots, N-1$) is first divided into L segments, each containing M data points. The ith segment of data is denoted as follows:

$$x_i(n) = x(n + iM - M), 0 \le n \le M, 1 \le i \le L. \quad (1)$$

Next, a window function $w(n)$ is applied to each data segment, and the periodogram for each segment is calculated. The periodogram for the ith segment is as follows:

$$\mathrm{I}_i(\omega) = \frac{1}{U}\left|\sum_{n=0}^{M-1} x_i(n)w(n)\mathrm{e}^{-j\omega n}\right|^2, i = 1, 2, 3, \ldots, M-1, \quad (2)$$

where the symbol U is the normalisation factor:

$$U = \frac{1}{M}\sum_{n=0}^{M-1} w^2(n), \quad (3)$$

If the periodograms of each segment are considered to be nearly uncorrelated, then the PSD estimate represented by P_{xx} is defined as follows:

$$\mathrm{P}_{xx}\left(e^{j\omega}\right) = \frac{1}{L}\sum_{i=1}^{L} \mathrm{I}_i(\omega). \quad (4)$$

Substituting $\omega = 2\pi f$ into the equation yields the following expression:

$$\mathrm{P}_{xx}(f) = \frac{1}{L}\sum_{i=1}^{L} \mathrm{I}_i(f). \quad (5)$$

Using the Welch method in Python, the $\mathrm{P}_{xx}(f)$ can be directly calculated, hence the PSD in the frequency band interval $[f_1 f_2]$ is as follows:

$$\mathrm{P}_{[f_1,f_2]} = \int_{f_1}^{f_2} \mathrm{P}_{xx}(f)\mathrm{d}f. \quad (6)$$

The relative PSD of the frequency band interval $[f_1,f_2]$ to the total frequency band interval $[f_0,f_3]$ can be obtained by division:

$$\mathrm{P} = \frac{\mathrm{P}_{[f_1,f_2]}}{\mathrm{P}_{[f_0,f_3]}}. \quad (7)$$

TABLE 1 Electroencephalography frequency bands and their characteristic.

Name	Frequency band	Characteristic
δ	0.3–5 Hz	Appearing during sleeping or unconscious states
θ	5–8 Hz	Appearing when dreaming or meditating
α	8–12 Hz	Appearing when relaxing
β_1	12–18 Hz	Appearing during a relaxed state of focused attention
β_2	18–30 Hz	Appearing when deeply concentrating
γ	30–45 Hz	Appearing during highly excited or stimulated states

3.3 | Feature fusion module

As illustrated in Figure 3, the feature fusion module includes a multimodal Transformer and a residual connection. The multimodal Transformer [13] is designed to enable one modality for

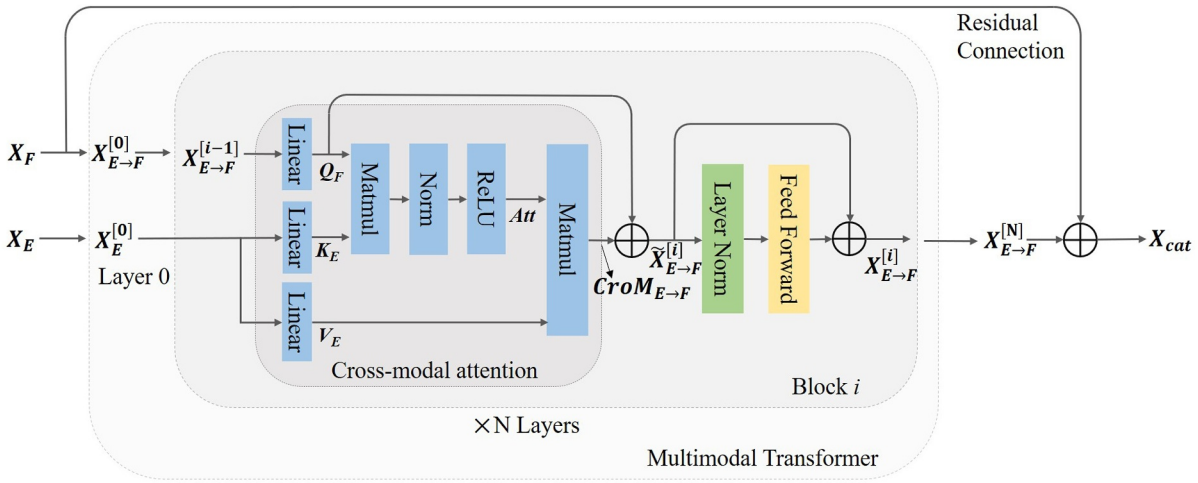


FIGURE 3 The illustration of the feature fusion module.

receiving information from another modality. Due to the various challenges and limitations of using EEG for emotion recognition, only one multimodal Transformer is employed to fully extract emotion-relevant features from EEG signals with the help of facial expressions features, which is denoted by '$E\to F$'.

Each multimodal Transformer is composed of N layer blocks. As for the ith layer ($i = 0,1, \ldots, N$), the cross-modal attention in multimodal Transformer is defined as follows:

$$\boldsymbol{X}_{E\to F}^{[0]} = \boldsymbol{X}_F, \tag{8}$$

$$\boldsymbol{Q}_F = \boldsymbol{X}_{E\to F}^{[i-1]} \boldsymbol{W}_Q^T, \tag{9}$$

$$\boldsymbol{K}_E = \boldsymbol{X}_E^{[0]} \boldsymbol{W}_K^T, \tag{10}$$

$$\boldsymbol{V}_E = \boldsymbol{X}_E^{[0]} \boldsymbol{W}_V^T, \tag{11}$$

$$\boldsymbol{Att} = \mathrm{ReLU}\left(\frac{\boldsymbol{Q}_F \boldsymbol{K}_E^T}{\sqrt{d_K}}\right), \tag{12}$$

$$\boldsymbol{CroM}_{E\to F} = \boldsymbol{Att}\boldsymbol{V}_E^T, \tag{13}$$

where $\boldsymbol{X}_E^{[0]}$ and $\boldsymbol{X}_F$ denote the EEG and face features respectively, $\boldsymbol{X}_{E\to F}^{[i-1]}$ is the output of the previous layer, $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V$ are the weights used to map the EEG signals onto the facial expression features. $\boldsymbol{Q}_F$ represents 'Query' which refers to the information to be queried. It determines what the model focuses on, indicating where to concentrate attention. $\boldsymbol{K}_E$ denotes 'Key', representing a set of reference data used to measure the similarity between the Query and other data. It reflects the inter-modality relationships. $\boldsymbol{V}_E$ is 'Value', representing the values associated with Key. Attention score ($\boldsymbol{Att}$) in Equation (12) is computed by measuring the similarity between $\boldsymbol{Q}_F$ and $\boldsymbol{K}_E$. To stabilise the gradient, the product of $\boldsymbol{Q}_F$ and $\boldsymbol{K}_E$ is normalised by dividing it by $\sqrt{d_K}$. Through the $\boldsymbol{Att}$, weights are assigned to the $\boldsymbol{V}_E$, and, ultimately, information from different modalities is aggregated through weighted summation into the output $\boldsymbol{CroM}_{E\to F}$. In comparison to the complexity of EEG, facial expressions are more intuitive. Additionally, results from experiments in single-modal emotion analysis also affirm that emotion-based approaches based on facial expressions often provide superior capabilities for capturing emotional information. Hence, in the first layer of the multimodal Transformer, facial expression signals are set as Query, serving as the reference for retrieval. Then features in the EEG modality that resemble facial expressions can be searched for and weights can be allocated to EEG features based on their similarity before they are integrated with facial features. Due to the strong association between facial expressions and emotions, the cross-modal attention mechanism allocates greater weight to EEG features associated with emotions, while assigning lower weight or even discarding redundant information.

After the EEG is weighted, the face features are supplemented and strengthened through residual connections in the cross-modal module, and then are normalised to obtain $\tilde{\boldsymbol{X}}_{E\to F}^{[i]}$:

$$\tilde{\boldsymbol{X}}_{E\to F}^{[i]} = \mathrm{LN}(\boldsymbol{Q}_F + CroM_{E\to F}), \tag{14}$$

where LN means layer normalisation.

Then, $\tilde{\boldsymbol{X}}_{E\to F}^{[i]}$ is fed into a feed-forward network:

$$\mathrm{F}\left(\tilde{\boldsymbol{X}}_{E\to F}^{[i]}\right) = \boldsymbol{W}_2\left(\mathrm{ReLU}\left(\boldsymbol{W}_1\tilde{\boldsymbol{X}}_{E\to F}^{[i]} + b_1\right)\right) + b_2, \tag{15}$$

where $\boldsymbol{W}_1$, b_1, $\boldsymbol{W}_2$ and b_2 are the parameters of the feedforward network.

Finally, the output of the ith layer of Transformer is defined as follows:

$$\boldsymbol{X}_{E\to F}^{[i]} = \tilde{\boldsymbol{X}}_{E\to F}^{[i]} + \mathrm{F}\left(\tilde{\boldsymbol{X}}_{E\to F}^{[i]}\right). \tag{16}$$

In the subsequent levels, the output will serve as a new Query for the continued exploration of emotion-related

features in the EEG modality, subsequently proceeding to another round of fusion. The number of repetitions in this process depends on the number of layers.

In order to prevent excessive loss of shallow features in a deep network, a residual connection [35] is incorporated into the fusion module to fuse the spatiotemporal features of shallow layers with the outputs of the multimodal Transformer. Ref. [32] indicates that facial expressions play a more important role in emotion analysis than EEG. Hence, only the facial expressions features are taken as an input of the residual connection to avoid redundancy:

$$\boldsymbol{X}_{cat} = \text{concat}\left(\boldsymbol{X}_F, \boldsymbol{X}_{E\to F}^{[N]}\right), \tag{17}$$

where N is the total layers of Transformer.

Finally, the concatenated $\boldsymbol{X}_{cat}$ is fed into a regressor for predicting the valence value.

3.4 | Loss function based on knowledge distillation

The loss function utilised in this study is the CCC loss. CCC is a widely used evaluation metric for regression problems, which can simultaneously measure the correlation and absolute difference between the predicted values $\boldsymbol{X} \in \mathrm{R}^{\mathrm{N}\times 1}$ and the true values $\boldsymbol{Y} \in \mathrm{R}^{\mathrm{N}\times 1}$. The formula is given as follows:

$$\rho_c = \frac{2\sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}, \tag{18}$$

where σ_{XY} is covariance, σ_X^2 and σ_Y^2 are variances, μ_X and μ_Y are means. The value of CCC ranges from -1 to 1, with a higher value indicating a stronger similarity between the predicted values and the true values, indicating a better predictive ability of the model. Thus, the CCC loss is defined as follows:

$$L_{\text{CCC}} = 1 - \rho_c. \tag{19}$$

Additionally, inspired by knowledge distillation, an improvement to the loss function is proposed in this paper.

Knowledge distillation is the process of transferring knowledge from a larger deep neural network to a smaller one and is commonly used for model compression. The larger one is named the teacher model while the smaller one is the student model. Cross-modal feature-level knowledge distillation is a technique to enhance model performance by transferring knowledge across different modalities, which is displayed in Figure 4a. Generally, the teacher model is trained on a more effective modality and then the outputs of a certain layer or several layers of it are saved. One of the main purposes of training the student model is to minimise the loss between its corresponding network layers outputs and those of the teacher model, as showed in Figure 4b. For example, Liu [36] trained a multispectral pedestrian detector (teacher model) on thermal images and then transferred the knowledge to a model that only receives RGB images, ultimately, they alleviated the reliance of existing multispectral pedestrian detectors on thermal images and the distillation method achieved robust performance.

In traditional knowledge distillation, the loss function is Kullback–Leibler divergence [37], while cross-entropy loss and L2 loss are also used in many papers. We utilise a sparser L1 loss, namely Knowledge distillation loss (KD loss), to measure the difference of given two feature vectors $\boldsymbol{U} \in \mathrm{R}^{\mathrm{T}\times\mathrm{H}}$ and $\boldsymbol{V} \in \mathrm{R}^{\mathrm{T}\times\mathrm{H}}$:

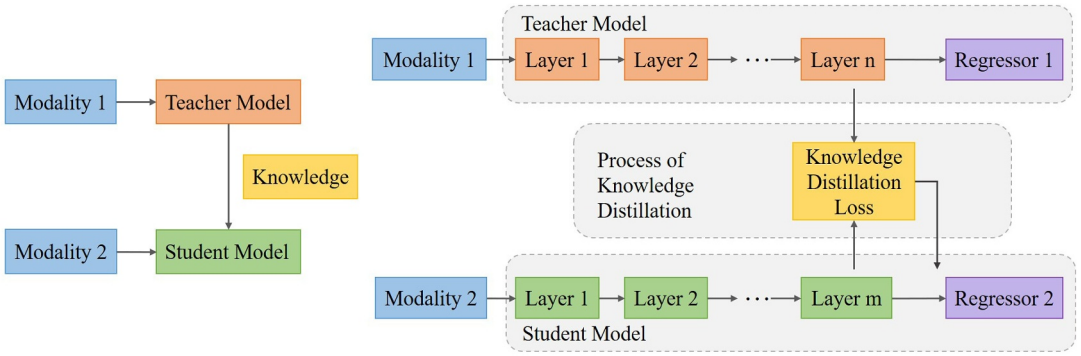
$$L_1(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{\mathrm{TH}} \sum_{i=1}^{\mathrm{T}} |u_i - v_i|, \tag{20}$$

where u_i and $v_i \in \mathrm{R}^{\mathrm{H}}$ are the feature points in each time step.

The loss function is introduced into the training of the multimodal fusion model and has a positive effect on model training which can be demonstrated by the experimental results. Therefore, the final loss function is defined as follows:

$$l = L_{\text{CCC}}(\boldsymbol{X}, \boldsymbol{Y}) + \alpha L_1(\boldsymbol{X}_E, \boldsymbol{X}_F), \tag{21}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the predicted and true values of valence respectively, $\boldsymbol{X}_E$ and $\boldsymbol{X}_F$ are EEG and facial expressions



(a) Simple diagram (b) Detailed diagram

FIGURE 4 Illustrations of cross-modal feature-level knowledge distillation.

features vectors respectively. The parameter α is used to control the degree of influence of KD loss on model training.

4 | EXPERIMENTS

4.1 | Dataset and data preprocessing

The MAHNOB-HCI dataset used in this paper was acquired by the Biosemi active II system with active electrodes. Thirty-two EEG electrodes were placed on a cap using international 10–20 system [38] as shown in Figure 5. Figure 5a is the EEG cap layout for 32 EEG electrodes and Figure 5b is the distribution of the four brain lobes, including the frontal, parietal, temporal and occipital lobes. Different colours indicate the correspondence between brain lobes and EEG electrodes.

To discuss the relationship between human emotions and each lobe of the brain. Data from several subjects with different emotional states and genders are selected. Table 2 illustrates their expressions and topo maps of their EEG PSD in six frequency bands during positive or negative emotions. In the topo maps, deeper colour means greater PSD and greater PSD means that part of the brain is more active.

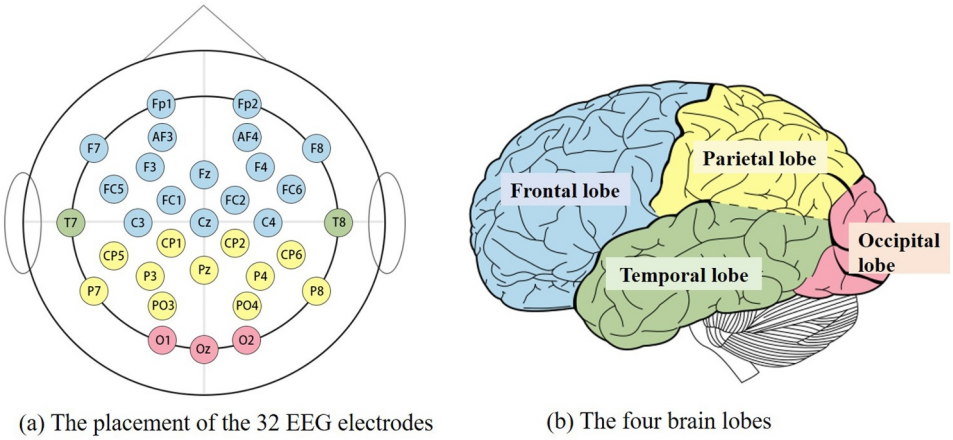
Taking into account all the maps, it is evident that all the four lobes can be active on the six bands. The reason is that during experiments, when the subjects are viewing stimulus videos, the occipital and temporal lobes perceive visual and auditory stimuli, the parietal lobe integrates perception, and the frontal lobe makes decisions and guides facial expressions. In fact, all parts of the brain should collaborate to produce emotions in real life.

Upon separate observation, it can be noted that in the β_1 band, β_2 band, and γ band, the colours are deeper on both sides of the brain. These three frequency bands tend to be associated with strong emotions, and thus human emotional activation channels are primarily distributed in the temporal lobe. In the δ band, the frontal lobe is the most active. The parietal lobes are the most active in the θ and α bands. Additionally, EEG is highly individualised. Comparing the first and the third subjects, they are both in the negative state, but their EEG topo maps are not the same. Therefore, how to train a common emotion recognition model for different people is also one of the future research directions.

The dataset is composed of facial expressions, audio signals, EEG, and peripheral physiological signals of 30 subjects when they were watching 20 videos, with video lengths ranging from 35 to 117 s. During each trial, subjects rate arousal, valence, dominance, and basic emotions categories with an integer scale ranging from 1 to 9. In addition to discrete emotion labels, a subset with continuous labels has been derived from the dataset. The subset consists of 239 experimental data from 24 subjects with clear facial expressions which are annotated by five experts using a joystick at 4 Hz. Since the arousal value is related to physical movements which cannot be observed solely through facial expression videos, only the valence is annotated. Valence values range from −0.5 to 0.5 and our experiments are carried out on this subset.

4.2 | Data preprocessing

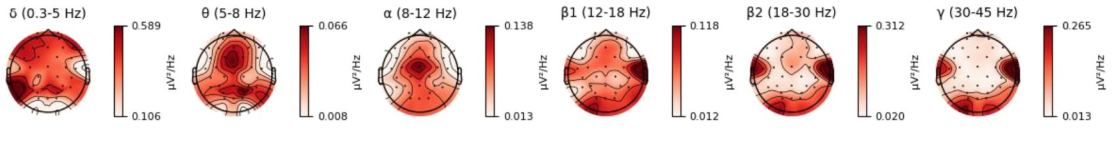
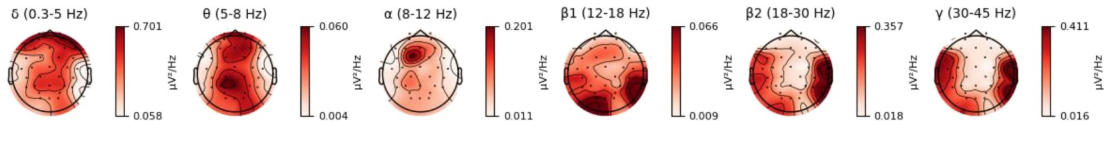
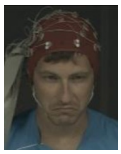
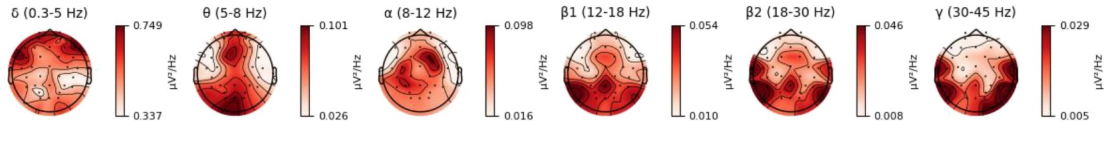
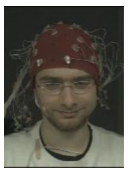
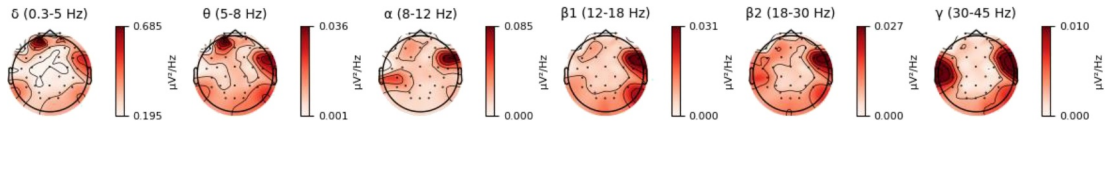
Preprocessing is required for both the facial expressions videos and the EEG data before they are fed to the model. For video data, the initial preprocessing step involves clipping the beginning and end of the video based on the timestamp information in the dataset, ensuring that only the reactions to emotional stimuli are preserved. Subsequently, the videos are changed to 64fps for synchronisation with the 4 Hz labels. Lastly, the video frames are cropped to a size of 48 × 48. For EEG data, the most crucial step is to remove artifacts with the average reference, which means each recorded EEG signal from any electrode needs to be subtracted by the average amplitude of all EEG signals from all electrodes. This is due to the fact that these average signals contain noise and artifacts that can be detected on the scalp but are not from the brain, such as eye movement signals, skin electrical signals etc. The introduction of an average reference can enhance the



(a) The placement of the 32 EEG electrodes (b) The four brain lobes

FIGURE 5 The illustration of the international 10–20 system and the brain lobes.

TABLE 2 PSD topo maps of subjects in the six frequency bands for different emotions.

Face	Topo maps of EEG PSD
	
	
	
	

Abbreviations: EEG, electroencephalography; PSD, power spectral density.

signal-to-noise ratio of EEG and consequently reduce errors in affective analysis.

4.3 | Data partitioning

The results are evaluated in a 10-fold cross-validation. In every fold, the dataset is divided into three sets, with 10% used for testing, 60% of the remaining samples used for training, and the rest used as validation sets. Specifically, the 239 trials are split into 129, 86, and 24 trials for training, validation, and testing in each fold. There may be data from the same subject in the training set, validation set, and test set, which may lead to better test results because the model has learnt similar data to the test set during training. So, it is neither subject-independent nor subject-dependent. To verify the generalisability of the RMMT across subjects, the leave-one-out cross-validation which is subject-independent is also employed. Specifically, the data from 24 subjects are divided into 24 folds according to the subjects, and one of the folds is used as the test set each time, while the data from the remaining 23 subjects are divided into the validation set and the test set in a ratio of 8:2.

4.4 | Evaluation metrics

The evaluation metrics are Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC), and CCC. Given the prediction $\boldsymbol{X} \in \mathrm{R}^{\mathrm{N}\times 1}$ and the continuous label $\boldsymbol{Y} \in \mathrm{R}^{\mathrm{N}\times 1}$, RMSE which measures the deviation between predicted and true values can be calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}(X_i - Y_i)^2}. \tag{22}$$

PCC and CCC are two commonly used correlation coefficients that reflect the degree of correlation between variables. PCC represents the linear correlation between variables and is calculated as following:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{23}$$

where $\boldsymbol{X}$ is the prediction and $\boldsymbol{Y}$ is the continuous label, σ_{XY} is the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and σ_X and σ_Y are the standard deviations of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. Values of PCC

range from −1 to 1, with 0 indicating no correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, a negative value indicating a negative correlation, and a positive value indicating a positive correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$. The closer the absolute value of PCC is to 1, the stronger the correlation between the two variables. CCC is an improvement over PCC as it not only considers the linearity between variables, but also focus on the numerical distance between them. The computation method is described in Formula (18). The value of CCC also falls within the range of −1 to 1, and the closer its absolute value is to 1, the closer the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

4.5 | Experimental results

Firstly, with the same feature extractors, two traditional fusion methods are used as the baseline models for the RMMT, including F-Fusion based on feature fusion, D-Fusion based on decision fusion. Numerical experiments have been conducted on the proposed RMMT and the two baseline models. Experimental results demonstrate that the proposed RMMT outperforms the benchmark models. Secondly, the RMMT is compared with previous methods to further prove the effectiveness of it. Eventually, ablation experiments about the residual module in the RMMT and KD loss function are performed to verify the necessity of them. The RMSE, PCC, and CCC in the experimental results are obtained by summing up the experimental results of each fold and then calculating the arithmetic mean in the 10-fold cross-validation or leave-one-out method.

4.5.1 | Models with different fusion methods

As mentioned earlier, there are various methods for multimodal fusion. To demonstrate the superior performance of the proposed method, a model based on feature fusion (F-Fusion) and a model based on decision fusion (D-Fusion) are designed. The model diagrams are illustrated in Figure 6, where $\boldsymbol{X}_F$ and $\boldsymbol{X}_E$ represent the spatiotemporal features extracted by the EEG and face feature extractors described in Chapter 3:

$$\boldsymbol{X}_F = [f_1\ f_2\ \ldots\ f_n], \boldsymbol{X}_E = [e_1\ e_2\ \ldots\ e_m]. \tag{24}$$

As shown in Figure 6a, in F-Fusion, $\boldsymbol{X}_F$ and $\boldsymbol{X}_E$ are directly concatenated to obtain $\boldsymbol{X}$ and then $\boldsymbol{X}$ is fed into the regression layer to predict the valence value:

$$\boldsymbol{X} = \text{concat}(\boldsymbol{X}_F, \boldsymbol{X}_E) = [f_1\ f_2\ \ldots\ f_n\ e_1\ e_2\ \ldots\ e_m]. \tag{25}$$

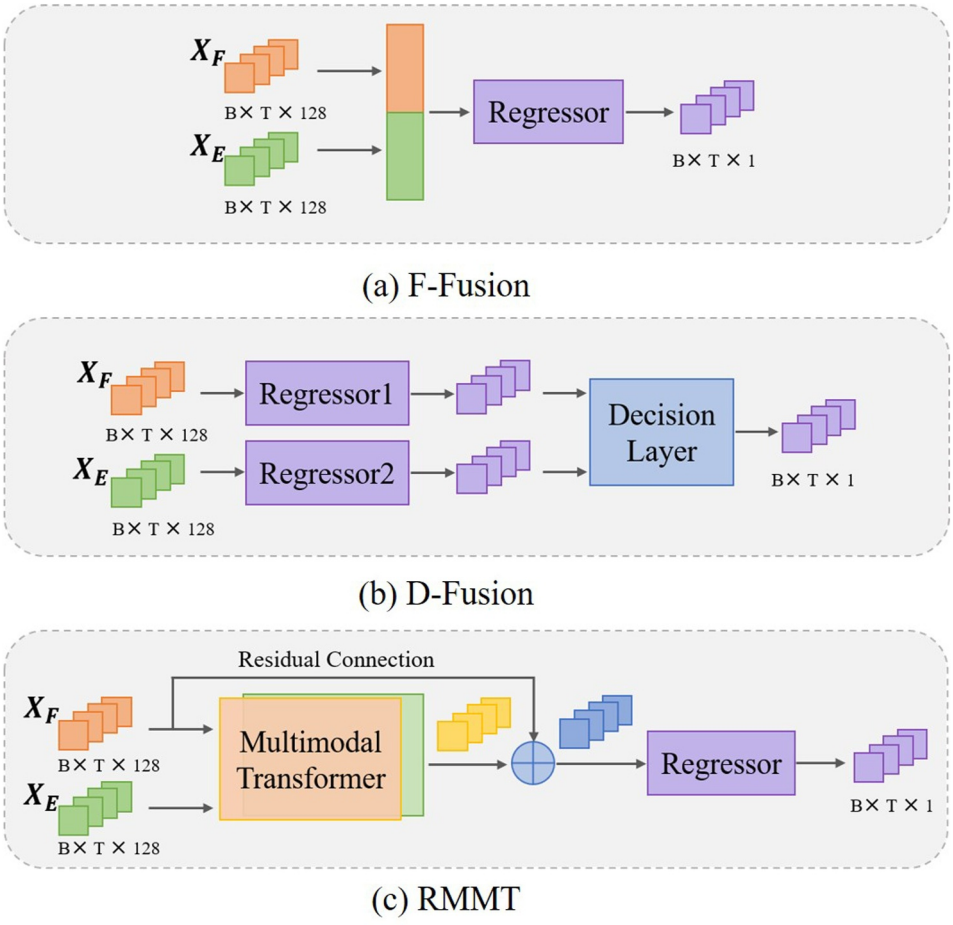


FIGURE 6 Illustrations of different fusion methods.

In Figure 6b, the D-fusion model takes $\boldsymbol{X}_F$ and $\boldsymbol{X}_E$ as inputs to their respective prediction layers. The predictions from both regressors are fused by means of a weighted average:

$$\boldsymbol{X} = \mu \boldsymbol{X}_F + \eta \boldsymbol{X}_E, \tag{26}$$

where μ and η represent the weights of facial expressions and EEG predictions, respectively. Different values are assigned to μ and η, and the experimental results are visualised in Figure 7. It reveals that the optimal values are achieved when $\mu = 0.6$ and $\eta = 0.4$, resulting in RMSE, PCC, and CCC values of 0.05, 0.747, and 0.713, respectively. Hence, $\mu = 0.6$ and $\eta = 0.4$ are assigned as the final values.

Figure 6c depicts a simplified diagram of the RMMT proposed in Chapter 3.

The results of each model are illustrated in Table 3. The fusion of EEG and facial expressions in F-Fusion and D-Fusion are relatively simple, leading to insufficient fusion, while the multimodal Transformer in the RMMT can consider contextual information and fully leverage the temporal relationships and spatiotemporal context among modalities during the fusion process, thereby enhancing the understanding and representation capability to data of the model. Hence the performance of the RMMT is optimal. While the RMSE values of these three models are comparable, the RMMT exhibits significantly better performance with PCC of 0.755, surpassing F-Fusion and D-Fusion by 0.017 and 0.008. The RMMT also achieves an optimum of 0.736 on the CCC with 4% and 3% more than F-Fusion and D-Fusion.

4.5.2 | Comparison with the state-of-the-art methods

The experimental results of the proposed RMMT as well as previous methods are displayed in Table 4. The results of refs. [29, 31] in the table are from the original paper, while the results of paper [32] are obtained by using the same parameters as our proposed model. They all use the 10-fold cross-validation.

On the one hand, for the single-modal experiments, the single-modal Transformers are added to the EEG feature extractor and the video feature extractor, respectively, obtaining residual single-modal Transformer for EEG (RSMT-E) and residual single-modal Transformer for face (RSMT-F). In the self-attention mechanism of the single-modal Transformer, each position is allowed to interact directly with all other positions, enabling the capture of long-range dependencies and thereby enhancing the accuracy. So, the obtained results of RSMT-E and RSMT-F are greater than the other models. In particular, the CCC value of RSMT-F is improved from 0.67 in ref. [32] to 0.701.

On the other hand, the multimodal experiments, the multimodal Transformer effectively enhances the fusion performance of multimodal data. As shown in the last row of Table 4, despite a slight increase in RMSE compared to the previous works, the PCC values of our model have significantly improved. The PCC values are improved by 65% and 42% compared to refs. [29, 31]. The main reason is the application of the multimodal Transformer with a residual connection module greatly improves the fusion effect. Multimodal Transformer facilitates interaction and information exchange between different modalities. Moreover, by utilising attention mechanism, the model captures the correlations and dependencies among different modalities. Another reason is that the feature extraction network for each modality demonstrates better effectiveness than refs. [29, 31]. For example, ref. [29] utilise traditional facial landmark features as video modality features,

TABLE 3 Results of different fusion methods.

	(a) F-Fusion	(b) D-Fusion	(c) RMMT
RMSE↓	0.050 ± 0.010	**0.050 ± 0.009**	0.051 ± 0.007
PCC↑	0.738 ± 0.082	0.747 ± 0.074	**0.755 ± 0.082**
CCC↑	0.706 ± 0.080	0.713 ± 0.076	**0.736 ± 0.078**

Note: ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

Abbreviations: CCC, consistency correlation coefficient; PCC, Pearson correlation coefficient; RMMT, residual multimodal Transformer; RMSE, root mean square error.

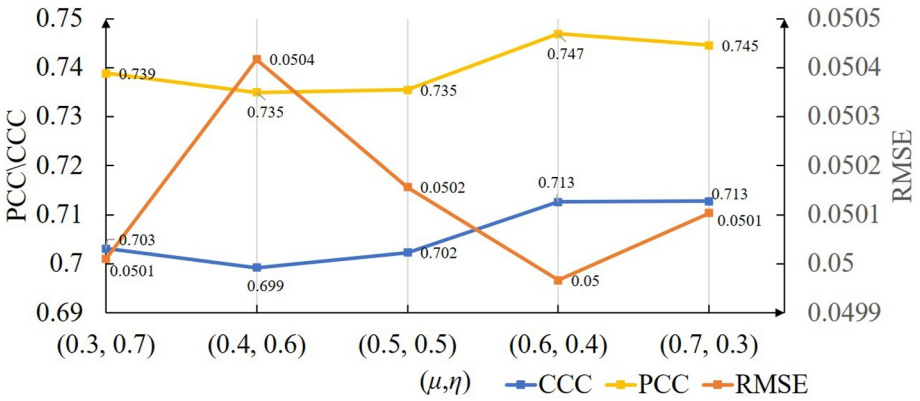


FIGURE 7 Results for different μ and η. CCC, consistency correlation coefficient; PCC, Pearson correlation coefficient; RMSE, root mean square error.

whereas more effective deep learning techniques are employed for feature extraction in our proposed model. And the TCN used for extracting temporal sequence features in the RMMT is also superior to the LSTM employed in ref. [29, 31].

The valence results of each single-modal and the RMMT of 10-fold cross-validation are separately plotted in Figure 8, as well as the target valence values to visualise the effectiveness of the model in emotion prediction and the performance of multimodal fusion. Results from two subjects are presented, respectively. From Figure 8a, it is evident that the curves of RSMT-F and RMMT closely match the target curve, indicating their effective capability in recognising emotional states. Moreover, the curves demonstrate consistency and stability in the RMMT for emotion analysis. For instance, in Figure 8a, the facial expressions at the 7th and 37th seconds are similar, and the corresponding valence values of the RMMT at these time points are also similar, with target values and the results of the RMMT are all around 0.25. The data at the 50th second reveals that results of the RMMT is closer to the target values, demonstrating the beneficial effect of EEG-face modality fusion in emotion analysis. Additionally, Figure 8b demonstrates negative emotions can also be recognised by the RMMT, despite slight deviations from the target values. The reason is that frowning amplitude in negative emotions is too small that it is more difficult to recognise than smiling in positive emotions, which is the same as ref. [29].

Table 5 presents the results from leave-one-out cross-validation, with the results of ref. [32] reproduced under the same configuration as the RMMT. Overall, the RSMT-F performs an improvement over ref. [32] and the RMMT outperforms unimodal models. However, the performance of the RSMT-E is slightly suboptimal compared to ref. [32]. This is attributed to the inherent individual differences in EEG, which result in a comparatively lower generalisation capability of

TABLE 4 Results from different models of the 10-fold cross-validation: '-' indicates missing data in the original paper.

	Models	RMSE↓	PCC↑	CCC↑
EEG	Soleymani [29]	0.053 ± 0.029	0.240 ± 0.340	-
	Choi [31]	0.049 ± 0.005	0.290 ± 0.080	-
	Zhang [32]	0.068 ± 0.017	0.492 ± 0.142	0.443 ± 0.148
	RSMT-E	0.077 ± 0.014	0.492 ± 0.135	0.450 ± 0.127
Face	Soleymani [29]	0.043 ± 0.026	0.480 ± 0.370	-
	Choi [31]	0.039 ± 0.004	0.520 ± 0.070	-
	Zhang [32]	0.056 ± 0.008	0.715 ± 0.079	0.670 ± 0.076
	RSMT-F	0.054 ± 0.009	0.727 ± 0.084	0.701 ± 0.079
Multimodal	Soleymani [29]	0.044 ± 0.026	0.450 ± 0.35	-
	Choi [31]	**0.037 ± 0.003**	0.530 ± 0.05	-
	Zhang [32]	0.068 ± 0.019	0.471 ± 0.155	0.42 ± 0.116
	RMMT	0.051 ± 0.007	**0.755 ± 0.082**	**0.736 ± 0.078**

Note: ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

Abbreviations: CCC, consistency correlation coefficient; PCC, Pearson correlation coefficient; RMMT, residual multimodal Transformer; RMSE, root mean square error; RSMT-E, residual single-modal Transformer for EEG; RSMT-F, residual single-modal Transformer for face.

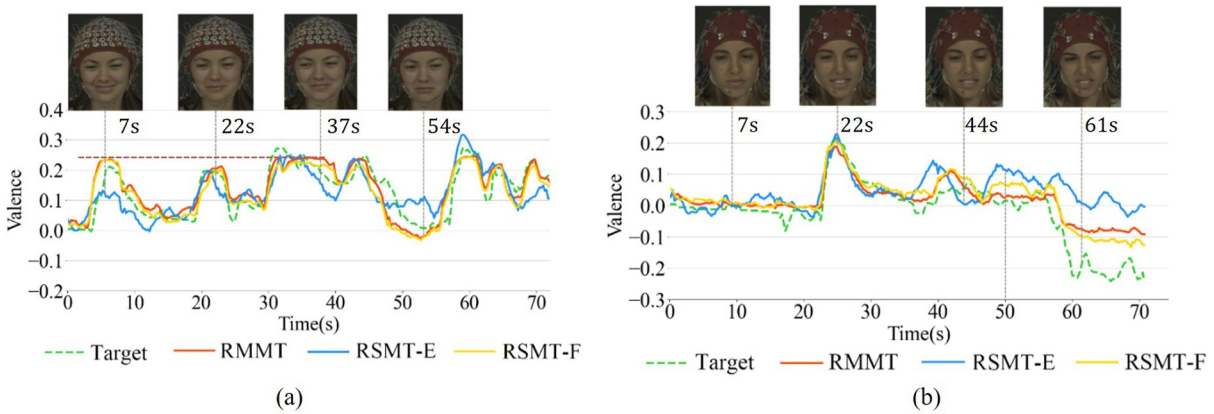


FIGURE 8 Valance results of proposed method plotting of the MAHNOB-HCI dataset. RMMT, residual multimodal Transformer; RSMT-E, residual single-modal Transformer for EEG; RSMT-F, residual single-modal Transformer for face.

emotion analysis models. Moreover, the limited size of EEG data is insufficient for training larger models, and larger models may introduce some noise, resulting in a decline in performance. In comparison with the results in Table 4 of the 10-fold cross-validation, the PCC and CCC of the RMMT drop by 7% and 14%. Better results in Table 4 are attributed to data leakage in 10-fold cross-validation, where training and testing sets may share data from the same subjects, allowing the model to prematurely learn similar features. Nevertheless, even using leave-one-out cross-validation, the performance of the RMMT remains superior to unimodal models from refs. [29, 31].

4.5.3 | Ablation experiments

Two ablation experiments are conducted in this section to further validate the performance of the RMMT, and the results are shown in Table 6.

Firstly, in order to verify the influences of the residual connection in the RMMT on the model performance, the residual connection is removed, obtaining the multimodal Transformer named MMT, and it is trained and tested on the MAHNOB-HCI dataset in the same way as the RMMT. The results are shown in the first column of Table 6. Theoretically, the spatiotemporal features of expressions extracted by ResNet50 and TCN may experience some information loss after repeated mapping and calculation in the multimodal Transformer. Because facial expression signals contribute more to emotion analysis results than EEG, the loss of facial expression information has a more negative impact on the results. The residual connection is employed to avoid this negative effect. As shown in Figure 6c, the residual connection enables the shallow facial expression features ($\boldsymbol{X_F}$) to skip the Transformer and directly input into the regression layer along with the output of the Transformer to predict the emotion state, and the $\boldsymbol{X_F}$ is utilised again to offset the loss caused by the Transformer, resulting in better experimental results. In multimodal fusion, if the contribution of different modalities is unbalanced, this approach can be considered to retain more effective information. From the experimental results, a comparison with the results of the RMMT reveals that the addition of the residual connection not only reduces the RMSE but also improves the PCC and CCC by 0.024 and 0.057, respectively, which means there is some crucial information lost in the multimodal Transformer.

In addition, to verify the impact of the feature-level KD loss, the RMMT is also trained without the KD loss. Results are shown in the second column of Table 6. Comparison with the RMMT shows that with the addition of KD loss, the RMSE is lower and both PCC and CCC are improved. The results suggest that the facial features can also effectively supervise the training of the model in multimodal fusion through KD loss.

Meanwhile, we conduct experiments to compare and select values of the weight α in the loss function. The value of α is set between 0.2 and 2, with an increment of 0.2. The variation of the results with the value of α is shown in Figure 9, and the details are shown in Table 7. When $\alpha < 1$, the impact of KD loss on model training is smaller than that of CCC loss, while it is opposite when $\alpha > 1$. It is evident from Figure 9 that the model can achieve favourable results when $\alpha \leq 1$, while the best performance is attained when $\alpha = 1$. As α increases beyond 1, both PCC and CCC gradually decrease. Obviously, this is because the influence of the KD loss becomes greater than that of the original labels, whereas the supervisory role of original labels is the most direct and effective in the deep learning model training process. The introduction of KD loss only helps the model extract more accurate features, but it cannot directly guide the model to predict emotional states.

TABLE 5 Results from different models of the leave-one-out method: '-' indicates missing data in the original paper.

	Models	RMSE↓	PCC↑	CCC↑
EEG	Zhang [32]	0.064 ± 0.026	0.498 ± 0.245	0.391 ± 0.245
	RSMT-E	0.070 ± 0.020	0.438 ± 0.254	0.374 ± 0.246
Face	Zhang [32]	0.054 ± 0.018	0.668 ± 0.231	0.574 ± 0.226
	RSMT-F	0.052 ± 0.013	0.675 ± 0.249	0.603 ± 0.225
Multimodal	**RMMT**	**0.047 ± 0.011**	**0.703 ± 0.195**	**0.635 ± 0.195**

Note: ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

Abbreviations: CCC, consistency correlation coefficient; PCC, Pearson correlation coefficient; RMMT, residual multimodal Transformer; RMSE, root mean square error; RSMT-E, residual single-modal Transformer for EEG; RSMT-F, residual single-modal Transformer for face.

TABLE 6 Results of ablation experiments.

	MMT	RMMT (without KD loss)	RMMT
RMSE↓	0.071 ± 0.012	0.053 ± 0.007	**0.051 ± 0.007**
PCC↑	0.731 ± 0.091	0.746 ± 0.087	**0.755 ± 0.082**
CCC↑	0.659 ± 0.098	0.722 ± 0.082	**0.736 ± 0.078**

Note: ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

Abbreviations: CCC, consistency correlation coefficient; MMT, multimodal Transformer; PCC, Pearson correlation coefficient; RMMT, residual multimodal Transformer; RMSE, root mean square error.

5 | CONCLUSION

In this paper, a new network for emotion recognition based on EEG and facial expressions named RMMT is proposed. Specifically, a face feature extractor and an EEG feature extractor are utilised to extract the spatiotemporal features of the two modalities separately. And then a multimodal Transformer is utilised to fuse the spatiotemporal features, and a residual connection is employed to compensate for the information loss caused by the deep neural network. The RMMT is verified on MANHOB-HCI dataset and the RMSE, PCC and CCC reach 0.051, 0.746 and 0.722 respectively, outperforming other state-of-the-art approaches.

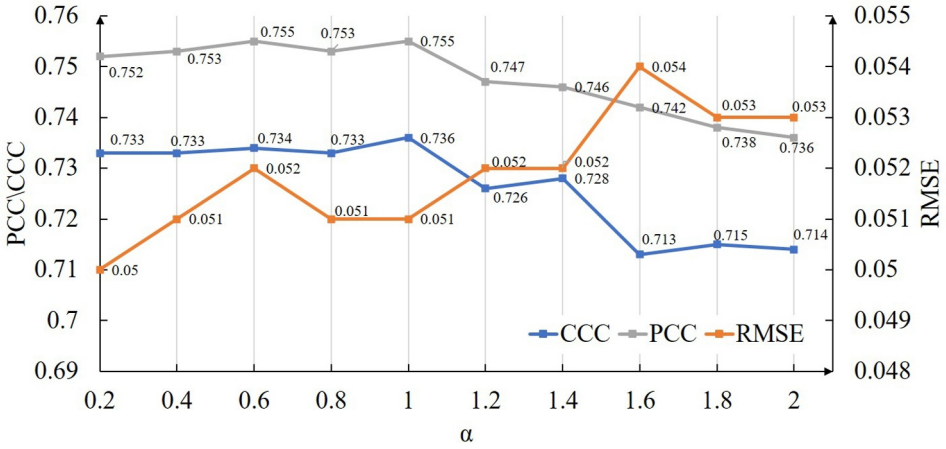


FIGURE 9 The variation of the results with the value of α. CCC, consistency correlation coefficient; PCC, Pearson correlation coefficient; RMSE, root mean square error.

TABLE 7 Specific results of different values of α.

α	0.2	0.4	0.6	0.8	1
RMSE↓	0.050 ± 0.007	0.051 ± 0.006	0.052 ± 0.007	0.051 ± 0.007	**0.051 ± 0.007**
PCC↑	0.752 ± 0.089	0.753 ± 0.077	0.755 ± 0.080	0.753 ± 0.082	**0.755 ± 0.082**
CCC↑	0.733 ± 0.091	0.733 ± 0.077	0.734 ± 0.075	0.733 ± 0.078	**0.736 ± 0.078**
α	1.2	1.4	1.6	1.8	2
RMSE↓	0.052 ± 0.008	0.052 ± 0.007	0.054 ± 0.009	0.053 ± 0.008	0.053 ± 0.008
PCC↑	0.747 ± 0.091	0.746 ± 0.088	0.742 ± 0.084	0.738 ± 0.082	0.736 ± 0.082
CCC↑	0.726 ± 0.085	0.728 ± 0.081	0.713 ± 0.087	0.715 ± 0.078	0.714 ± 0.073

Note: ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

Abbreviations: CCC, consistency correlation coefficient; PCC, Pearson correlation coefficient; RMSE, root mean square error.

Due to the fact that difficulty in recognising frowns in negative emotions due to their small amplitude, the performance of the RMMT in negative emotions is slightly weak. This is also one of the directions for future improvement of our model. In addition, it can be seen from the residual connection structure and loss function of the model that the RMMT relies too heavily on the guidance of facial expressions in extracting emotional information from EEG during multimodal fusion. But theoretically, EEG contains very abundant and available information that can be utilised for better experimental results. Therefore, in subsequent studies, more features can be considered to be extracted and utilised from EEG, such as statistical features in the time domain, rational asymmetric (RASM) features in the frequency domain, and wavelet transform features in the time-frequency domain. This may reduce the reliance on facial expressions and address the imbalance in emotion recognition based on the two modalities. Furthermore, RMMT currently utilises only EEG and facial expressions as input. In the future, the incorporation of additional physiological signals, such as electrocardiogram, electromyogram, and galvanic skin response, could be explored.

ACKNOWLEDGEMENTS

This study was funded by the State Key Development Program in 14th Five-Year under Grant No. 2021YFF0900701.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from Imperial College London. Restrictions apply to the availability of these data, which were used under licence for this study. Data are available at https://mahnob-db.eu/hci-tagging/ with the permission of Imperial College London.

ORCID

Jieyu Xiao https://orcid.org/0009-0005-2752-4597

REFERENCES

1. Ekman, P.: An argument for basic emotions. Cognit. Emot. 6(3-4), 169–200 (1992). https://doi.org/10.1080/02699939208411068
2. Lang, P.J.: The emotion probe: studies of motivation and attention. Am. Psychol. 50(5), 372–385 (1995). https://doi.org/10.1037//0003-066x.50.5.372

3. Jin, X., Lai, Z., Jin, Z.: Learning dynamic relationships for facial expression recognition based on graph convolutional network. IEEE Trans. Image Process. 30, 7143–7155 (2021). https://doi.org/10.1109/tip.2021.3101820
4. Zhi, R., et al.: Action unit analysis enhanced facial expression recognition by deep neural network evolution. Neurocomputing 425, 135–148 (2021). https://doi.org/10.1016/j.neucom.2020.03.036
5. Wang, H., et al.: Emotion expression with fact transfer for video description. IEEE Trans. Multimed. 24, 715–727 (2021). https://doi.org/10.1109/tmm.2021.3058555
6. Lew, W.-C.L., et al.: EEG-video emotion-based summarization: learning with EEG auxiliary signals. IEEE Trans. Affect. Comput. 13(4), 1827–1839 (2022). https://doi.org/10.1109/taffc.2022.3208259
7. Du, C., et al.: Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. IEEE Trans. Pattern Anal. Mach. Intell. 45(9), 10760–10777 (2023). https://doi.org/10.1109/tpami.2023.3263181
8. Liu, Y., Fu, G.: Emotion recognition by deeply learned multi-channel textual and EEG features. Future Generat. Comput. Syst. 119, 1–6 (2021). https://doi.org/10.1016/j.future.2021.01.010
9. Fang, Y., et al.: Multi-feature input deep forest for EEG-based emotion recognition. Front. Neurorob. 14, 617531 (2021). https://doi.org/10.3389/fnbot.2020.617531
10. Kang, J.S., et al.: ICA-evolution based data augmentation with ensemble deep neural networks using time and frequency kernels for emotion recognition from EEG-data. IEEE Trans. Affect. Comput. 13(02), 616–627 (2022). https://doi.org/10.1109/taffc.2019.2942587
11. Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41(2), 423–443 (2018). https://doi.org/10.1109/tpami.2018.2798607
12. Yin, G., et al.: An efficient multimodal framework for large scale emotion recognition by fusing music and electrodermal activity signals. arXiv preprint arXiv:2008.09743 (2020)
13. Wu, Di, Zhang, J., Zhao, Q.: Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning. IEEE Access 8, 133180–133189 (2020). https://doi.org/10.1109/access.2020.3010311
14. Tsai, Y.-H.H., et al.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference. Association for Computational Linguistics, vol. 2019. NIH Public Access (2019). https://doi.org/10.18653/v1/p19-1656
15. Rahimpour, M., et al.: Cross-modal distillation to improve MRI-based brain tumor segmentation with missing MRI sequences. IEEE Trans. Biomed. Eng. 69(7), 2153–2164 (2022). https://doi.org/10.1109/tbme.2021.3137561
16. Chen, L., et al.: Electroglottograph-based speech emotion recognition via cross-modal distillation. Appl. Sci. 12(9), 4338 (2022). https://doi.org/10.3390/app12094338
17. Yuan, Z., et al.: A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. IEEE Trans. Geosci. Rem. Sens. 60, 3124252 (2022). https://doi.org/10.1109/tgrs.2021.3124252
18. D'mello, S.K., Kory, J.: A review and meta-analysis of multimodal affect detection systems. ACM Comput. Surv. 47(3), 1–36 (2015). https://doi.org/10.1145/2682899
19. Du, C., et al.: Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In: Proceedings of the 26th ACM International Conference on Multimedia (2018)
20. Sun, X., et al.: Multimodal emotion classification method and analysis of brain functional connectivity networks. IEEE Trans. Neural Syst. Rehabil. Eng. 30, 2022–2031 (2022). https://doi.org/10.1109/tnsre.2022.3192533
21. Tian, J., She, Y.: A visual–audio-based emotion recognition system integrating dimensional analysis. IEEE Trans. Comput. Soc. Syst. 10(6), 3273–3282 (2022). https://doi.org/10.1109/tcss.2022.3200060
22. Ayari, N., et al.: Hybrid model-based emotion contextual recognition for cognitive assistance services. IEEE Trans. Cybern. 52(5), 3567–3576 (2022). https://doi.org/10.1109/tcyb.2020.3013112
23. Zhou, H., et al.: Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 2617–2629 (2021). https://doi.org/10.1109/taslp.2021.3096037
24. Wang, Z., et al.: Spatial-temporal feature fusion neural network for EEG-based emotion recognition. IEEE Trans. Instrum. Meas. 71, 1–12 (2022). https://doi.org/10.1109/tim.2022.3165280
25. Pei, E., et al.: Monocular 3D facial expression features for continuous affect recognition. IEEE Trans. Multimed. 23, 3540–3550 (2020). https://doi.org/10.1109/tmm.2020.3026894
26. Chen, H., Jiang, D., Sahli, H.: Transformer encoder with multi-modal multi-head attention for continuous affect recognition. IEEE Trans. Multimed. 23, 4171–4183 (2020). https://doi.org/10.1109/tmm.2020.3037496
27. Hu, M., et al.: A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video. IEEE Signal Process. Lett. 28, 698–702 (2021). https://doi.org/10.1109/lsp.2021.3063609
28. Li, X., et al.: A multi-scale multi-task learning model for continuous dimensional emotion recognition from audio. Electronics 11(3), 417 (2022). https://doi.org/10.3390/electronics11030417
29. Soleymani, M., et al.: Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Trans. Affect. Comput. 7(1), 17–28 (2016). https://doi.org/10.1109/taffc.2015.2436926
30. Li, D., et al.: The fusion of electroencephalography and facial expression for continuous emotion recognition. IEEE Access 7, 155724–155736 (2019). https://doi.org/10.1109/access.2019.2949707
31. Choi, D.Y., Kim, D.-H., Song, B.C.: Multimodal attention network for continuous-time emotion recognition using video and EEG signals. IEEE Access 8, 203814–203826 (2020). https://doi.org/10.1109/access.2020.3036877
32. Zhang, Su, Tang, C., Guan, C.: Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition. Pattern Recogn. 130, 108833 (2022). https://doi.org/10.1016/j.patcog.2022.108833
33. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
34. Bai, S., Zico Kolter, J., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
35. Tao, H., et al.: Multi-stream convolution-recurrent neural networks based on attention mechanism fusion for speech emotion recognition. Entropy 24(8), 1025 (2022). https://doi.org/10.3390/e24081025
36. Liu, T., et al.: Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. IEEE Trans. Circ. Syst. Video Technol. 32(1), 315–329 (2022). https://doi.org/10.1109/tcsvt.2021.3060162
37. Song, J., et al.: Spot-adaptive knowledge distillation. IEEE Trans. Image Process. 31, 3359–3370 (2022). https://doi.org/10.1109/tip.2022.3170728
38. Homan, R.W.: The 10–20 electrode system and cerebral location. Am. J. EEG Technol. 28(4), 269–279 (1988). https://doi.org/10.1080/00029238.1988.11080272

How to cite this article: Jin, X., et al.: Residual multimodal Transformer for expression-EEG fusion continuous emotion recognition. CAAI Trans. Intell. Technol. 9(5), 1290–1304 (2024). https://doi.org/10.1049/cit2.12346