frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 10 January 2024
DOI 10.3389/fnins.2023.1330077

OPEN ACCESS

EDITED BY
Benjamin Thompson,
University of Waterloo, Canada

REVIEWED BY
Junsong Wang,
Shenzhen Technology University, China
Weicheng Xie,
Shenzhen University, China

*CORRESPONDENCE
Penghai Li
✉ lph1973@tju.edu.cn
Longlong Cheng
✉ chenglonglong@cecdat.com

RECEIVED 30 October 2023
ACCEPTED 19 December 2023
PUBLISHED 10 January 2024

CITATION
Du Y, Li P, Cheng L, Zhang X, Li M and Li F (2024) Attention-based 3D convolutional recurrent neural network model for multimodal emotion recognition. *Front. Neurosci.* 17:1330077. doi: 10.3389/fnins.2023.1330077

Attention-based 3D convolutional recurrent neural network model for multimodal emotion recognition

Yiming Du[1], Penghai Li[1]*, Longlong Cheng[1,2]*, Xuanwei Zhang[3], Mingji Li[1] and Fengzhou Li[1]

[1]School of Integrated Circuit Science and Engineering, Tianjin University of Technology, Tianjin, China, [2]China Electronics Cloud Brain (Tianjin) Technology Co, Ltd., Tianjin, China, [3]School of Information Engineering, China University of Geosciences, Beijing, China

Introduction: Multimodal emotion recognition has become a hot topic in human-computer interaction and intelligent healthcare fields. However, combining information from different human different modalities for emotion computation is still challenging.

Methods: In this paper, we propose a three-dimensional convolutional recurrent neural network model (referred to as 3FACRNN network) based on multimodal fusion and attention mechanism. The 3FACRNN network model consists of a visual network and an EEG network. The visual network is composed of a cascaded convolutional neural network–time convolutional network (CNN-TCN). In the EEG network, the 3D feature building module was added to integrate band information, spatial information and temporal information of the EEG signal, and the band attention and self-attention modules were added to the convolutional recurrent neural network (CRNN). The former explores the effect of different frequency bands on network recognition performance, while the latter is to obtain the intrinsic similarity of different EEG samples.

Results: To investigate the effect of different frequency bands on the experiment, we obtained the average attention mask for all subjects in different frequency bands. The distribution of the attention masks across the different frequency bands suggests that signals more relevant to human emotions may be active in the high frequency bands γ (31–50 Hz). Finally, we try to use the multi-task loss function Lc to force the approximation of the intermediate feature vectors of the visual and EEG modalities, with the aim of using the knowledge of the visual modalities to improve the performance of the EEG network model. The mean recognition accuracy and standard deviation of the proposed method on the two multimodal sentiment datasets DEAP and MAHNOB-HCI (arousal, valence) were 96.75 ± 1.75, 96.86 ± 1.33; 97.55 ± 1.51, 98.37 ± 1.07, better than those of the state-of-the-art multimodal recognition approaches.

Discussion: The experimental results show that starting from the multimodal information, the facial video frames and electroencephalogram (EEG) signals of the subjects are used as inputs to the emotion recognition network, which can enhance the stability of the emotion network and improve the recognition accuracy of the emotion network. In addition, in future work, we will try to utilize sparse matrix methods and deep convolutional networks to improve the performance of multimodal emotion networks.

KEYWORDS

electroencephalogram (EEG), emotion recognition, attention mechanism, convolutional neural network (CNN), 3D feature construction module, multimodal recognition

1 Introduction

Emotion recognition and analysis are crucial in our everyday lives, particularly in the fields of human-computer interaction (Liu Y. et al., 2020; Cheng et al., 2021), the assessment of psychological disorders such as depression and autism (Blankertz et al., 2016), and fatigue driving (Kong et al., 2017). There are two distinct categories of emotional recognition signals: physiological and non-physiological. Electromyography (EMG), electroencephalography (EEG), electrocardiogram, heart rate (Doma and Pirouz, 2020) and respiratory rate are examples of physiological signals, while facial expressions, utterances, and body postures are examples of non-physiological signals (Daros et al., 2013; Huang et al., 2020).

EEG is noninvasive, practical, quick, and affordable. Consequently, it is frequently employed to examine the brain's response to emotional stimuli. We can acquire emotion-related feature information from different frequency bands and electrodes of the EEG, and use deep learning methods for feature learning and classification. Wang et al. (2018) used a 3D CNN network to extract spatial features from EEG signals followed by emotion state prediction, but did not consider the effect of temporal feature components in EEG signals on emotion recognition; Yang et al. (2018a) extracted spatio-temporal feature information from EEG signals by cascading CNN and LSTM networks, which is similar to the emotion recognition architecture based on convolutional recurrent networks proposed in this paper, but the method proposed by Yang et al. did not integrate feature information of EEG data in different dimensions, which resulted in spatio-temporal features representativeness extracted by the CNN-LSTM network did not comprehensive; Li et al. (2018) constructed a two-dimensional matrix of 62 electrode locations and mapped the EEG features onto the two-dimensional matrix, they were then fed into a network model for training; Song et al. designed differential entropy features based on the relationship between electrode locations and used a graph convolutional neural network as a classifier (Song et al., 2020); both Li et al. and Song et al. only considered the effect of relative position information between different electrodes on emotion recognition, ignoring the importance of information from different frequency bands within the same electrode for the prediction of emotional states; Yang et al. used a combination of four frequency bands in the EEG, including theta (4–7 Hz), α (8–13 Hz), β (14–30 Hz), and γ (31–50 Hz), and found that they are closely related to emotional states (Yang et al., 2018b), but did not use attentional means to adjust the weight parameters of the different frequency bands according to their importance to help the emotion network to better fulfill the emotion prediction task. To address the weaknesses and shortcomings of the above research methods, in this paper, we propose a three-dimensional convolutional recurrent neural network model based on the attention mechanism, 3FACRNN, which can first integrate the multidimensional feature information of EEG signals using the three-dimensional feature construction module to increase the feature complexity of EEG signals, then extract the deep spatio-temporal features of EEG signals using the convolutional recurrent neural network, and finally combine with the frequency-band attention module and the self-attention module to improve the discriminative capability of the feature information.

Inspired by the research of Shen et al. (2020), this paper proposes a 3D feature construction module to better utilize all the emotional information contained in the EEG signals. This 3D feature building module can extract frequency band, spatial and temporal information from the original EEG signals, and then input the 3D-structured EEG signals into a neural network consisting of CNNs and LSTMs for deeper feature abstraction, and finally input them into a SoftMax classifier for emotional state classification. Incorporating attentional mechanisms, such as frequency bands and self-attention mechanisms into this procedure allows us to extract more discriminative feature information (Guo et al., 2022; Tao et al., 2023). Although all four bands of EEG signals contain information related to emotions, the importance of the emotional information contained in different bands varies. To deal with this case, we used 1×1 convolution method to assign different weights to different bands. In addition, since the importance of different EEG samples of subjects varies, we integrate a self-attention module in LSTM, which can extract the attention information of subjects according to the importance of their different EEG samples. Through the attention mechanism, the 3FACRNN network is able to acquire more discriminative feature information from EEG signals, thereby enhancing its recognition performance.

The EEG signal is the result of the integrated activity of human brain regions, and because it is not influenced by subjective human factors, it accurately reflects the true emotional state of a person in response to a stimulus. However, the EEG signal is easily disturbed by noise. Although facial expression can visually communicate the subject's emotional state, it is often disguised, so the subject is sometimes unable to express his or her own emotional state accurately. Multimodal emotion recognition methods that combine the facial expressions of subjects with EEG signals can compensate for the deficiencies of unimodal methods and achieve superior recognition results (Mühl et al., 2014; D'mello and Kory, 2015; Basbrain and Gan, 2020). Afouras et al. (2020) trained a visual recognition model based on lip reading using the knowledge of obscurity in the speech modality, but both the speech and visual modalities are artifactual and do not allow for true emotional state labeling. Soleymani et al. (2016) proposed a multimodal continuous emotion prediction method based on facial sign sequences and EEG signals, obtaining high recognition accuracy. Tzirakis et al. (2017) achieved the first end-to-end emotion recognition by using ResNet and two convolutional layers to extract facial expression feature signals and speech feature signals, concatenating them to form new features, and then integrating contextual information via a multilayer LSTM. Both the studies of Soleymani et al. and Tziraki et al. fused the feature information of the two modalities in series at the feature level or decision level, without considering the differences and complementarities between the features. In contrast, the multimodal emotional network 3FACRNN network proposed in this paper does not serially splice features between two modalities and feed them into the network for undifferentiated learning, but rather forces approximation of the intermediate feature vectors of the visual and EEG modalities through the multitasking loss function Lc, with the aim of utilizing the knowledge of the visual modalities for the improvement of the performance of the EEG network model.

In this paper, we propose a novel multimodal emotion recognition network (3FACRNN) based on the attention mechanism, which includes visual and EEG networks. The visual network is trained only for the visual modality, and the important feature

information in the visual modality is extracted and used to supervise the EEG network for training and learning. The proposed multi-task loss function L_C consists of the weighted sum of the L_1 loss function and the cross-entropy loss function, which is used to force the approximation of the intermediate feature vectors of the two modalities, so that the EEG modalities can learn the knowledge of the visual modalities, thereby improving the recognition performance of the EEG network. The EEG network model cascades the 3D feature construction module, the multi-attention module, the CNN, and the LSTM framework, The inputs to the EEG network include intermediate feature vectors obtained from the visual network and raw EEG signals and labels. The raw EEG signals are first preprocessed to remove noise, artifacts, and baseline signals, and then important feature information is integrated into the EEG signals using a 3D feature building module, and then important band information and intrinsic similarity information of different EEG signals are extracted using a multi-attention module, and finally, a convolutional recurrent neural network to extract local features for deeper feature abstraction, and a classifier consisting of a fully connected layer and SoftMax is used to complete the prediction of emotion labels. The proposed 3FACRNN network model for emotion recognition has been evaluated on two publicly available datasets, the DEAP dataset (Koelstra et al., 2012) and the MAHNOB-HCI dataset (Soleymani et al., 2012). On both datasets, the network model has demonstrated outstanding recognition accuracy. Here is a synopsis of our most notable contributions:

1 This paper proposes a 3D convolutional recurrent neural network model based on the attention mechanism called 3FACRNN, which cascades a 3D feature construction module, a frequency band attention module, a convolutional recurrent neural network, and a self-attention module to perform the emotion recognition task. This model can effectively enhance the discriminative properties of EEG signals in space, time, and spectrum.
2 In this paper, we use the multi-task loss function Lc to force approximation of the intermediate feature vectors of visual modality and EEG modality in order to achieve the purpose of using the dark knowledge of visual modality to supervise the emotion recognition of the EEG network, which effectively utilizes the advantage of the high resolution of the visual modality in spatial dimensions, solves the problem of the single data information in the uni-modal approach, and improves the feature complexity of the EEG signals in spatial dimensions.
3 The average accuracy and standard deviation of the proposed 3FACRNN model on the valence and arousal dimensions of the DEAP and MAHNOB-HCI datasets were 96.75 ± 1.75, 96.86 ± 1.33, 97.55 ± 1.51, and 98.37 ± 1.07. It outperforms existing emotion recognition methods using multimodal data. Moreover, this paper analyses the attentional weights of various frequency bands, and the weight distribution suggests that the gamma band of EEG signals may be more pertinent to human emotions.

The remaining sections of the paper are organized as follows: section 2 describes the relevant materials and methodologies, section 3 analyses the experimental results, section 4 discusses the work accomplished and concludes the entire paper.

2 Methods

The framework of the proposed 3FACRNN multimodal emotional network model is shown in Figure 1. It is made up of the visual and EEG networks.

The facial video frames of the subjects in the DEAP dataset and the MAHNOB-HCI dataset were then fed into the pre-trained visual network to extract the spatio-temporal eigenvectors of the visual modalities. The feature information obtained from the visual modalities along with the original EEG signal and labels was fed into the EEG network so that the dark knowledge of the visual modalities could enhance the recognition performance of the EEG network. The specific implementation process of each component in the two subnetworks and the interaction mechanism between the two subnetworks for learning will be described in detail in the following section.

2.1 The visual network model

The visual network model is shown in Figure 2. The visual network model consists of facial video frame acquisition, pre-processing, CNN, time series convolutional network (TCN) and SoftMax classifier architecture. The facial expression is first processed by frame extraction, and the processed face video frames are resized to 48×48 ×3, with T video frames input at a time. It is then passed through spatial–temporal convolutional network, which consists of CNN, TCN and linear layers. The T consecutive video frames and labels are fed into a CNN, which contains two convolutional layers, two pooling layers and a flatten layer. The spatial information of the video frames is derived by the CNN network and spatial features are generated for each frame, resulting in T×512-dimensional spatial features. The latter is then passed through a temporal convolutional neural network (TCN), from which temporal information is obtained and the spatial–temporal composite features of the video frames are obtained, generating T×128-dimensional spatial–temporal features. TCN networks are capable of extracting features at different time scales and can effectively capture long-term dependencies in time-series data (Xue et al., 2020; He et al., 2022; Wang et al., 2022). The TCN network in the visual network proposed in this paper cascades two temporal convolution modules and a pooling layer, and the internal parameters of both temporal convolution modules can be shared. The temporal convolution module consists of one normal convolutional layer, two dilated convolutional layers, and a residual block, with Relu activation functions and normalization layer added after each convolutional and dilated convolutional layer. The convolution kernel size of the convolutional layer is 3, stride = 1, dilation = 1; the convolution kernel size of the dilated convolutional layer is 3, stride = 1, dilation = 2. The role of the dilated convolutional layer is to inject voids into the convolutional layer as a way to increase the receptive field so that the output contains a larger range of feature information than it otherwise would. Where the dilation parameter refers to the number of kernel intervals. A 1×1 convolutional kernel is used in the residual block to perform dimensional matching of the input to the output and to residually connect the input to the output to prevent gradient explosion. In addition we added Dropout layer after each normalization layer for preventing overfitting and its scale is set to 0.5. The $T \times 128$ dimensional spatio-temporal features are then

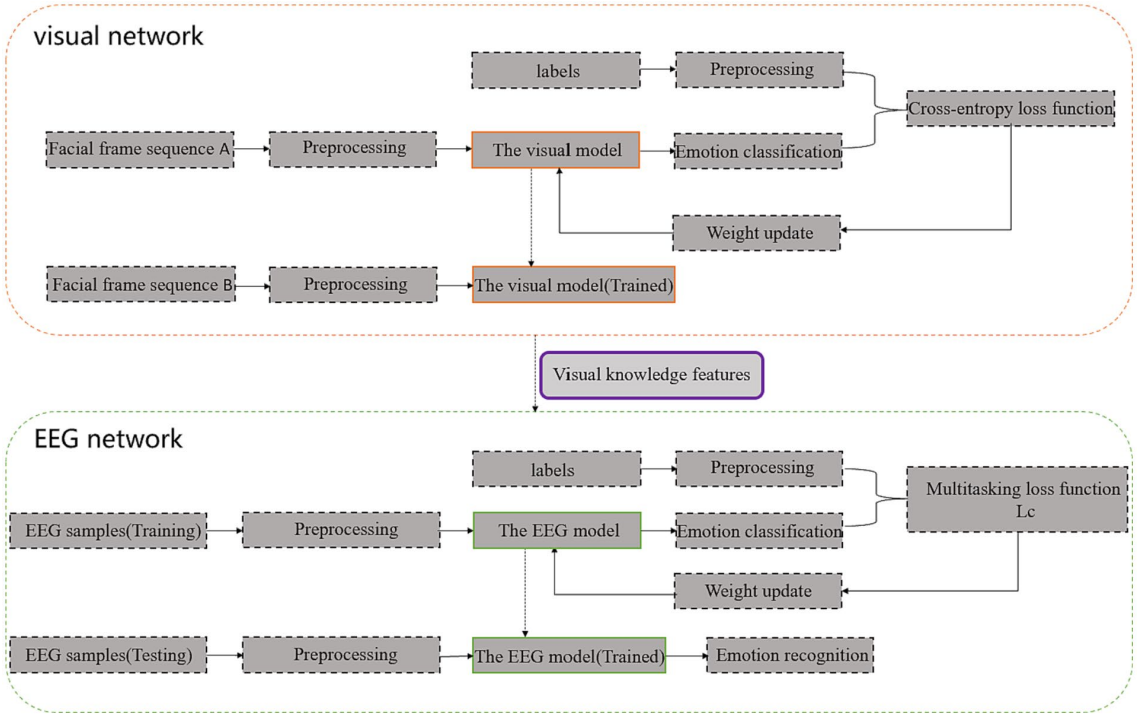

FIGURE 1
3FACRNN network architecture (Facial frame sequence A is obtained after preprocessing of facial expression videos in the AFEW-VA dataset, and f Facial frame sequence B is obtained after preprocessing of facial expression videos in the DEAP dataset and MAHNOB-HCI dataset).

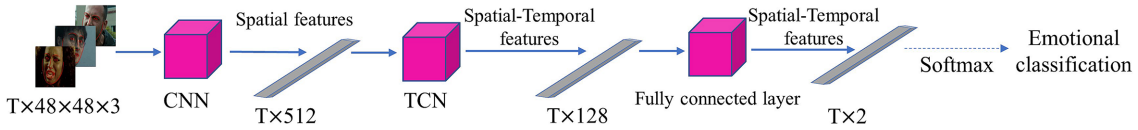

FIGURE 2
Framework for the visual model.

mapped to T × 2 dimensions using a fully connected layer. Finally, the SoftMax classifier receives the extracted features as input to recognize the emotional state., using a cross-entropy loss function to reduce the distance between the predicted sequence and the actual sequence (labels). We used a large number of 2D-based facial image or emotion databases AFEW-VA (Kossaifi et al., 2017) and AffectNet (Mollahosseini et al., 2019) to train the visual network.

2.2 The EEG network model

Figure 3 depicts the EEG network model, which consists of an EEG signal preprocessing, a 3D feature construction module (feature extraction), a convolutional recurrent neural network (frequency band attention module, CNN, LSTM, self-attention module), and a SoftMax classifier. First, we pre-process the raw EEG signal to eliminate the baseline signal, and then we input the pre-processed EEG signal into the 3D feature construction module to integrate the frequency information, spatial information, and temporal information of the signal. The 3D EEG signals are then fed into a convolutional recurrent neural network. The frequency band attention module in the convolutional recurrent neural network captures the frequency bands that are more critical to the task. The self-attention mechanism focuses on more important EEG samples by assessing the probability based on the similarities between samples. The CNN and LSTM networks further abstract and extract the spatial–temporal features of the EEG signals. Finally, a SoftMax classifier is utilized to predict the subject's emotional state. Each component's implementation is described in detail below.

2.2.1 3D feature construction module

Initially, we carry out preprocessing procedures on the unprocessed EEG signals, which encompass both baseline and experimental signals (Ahmed et al., 2023). In this paper, we use a non-overlapping sliding window to remove the baseline signal from the raw EEG signal. According to previous studies, human emotional state is generally maintained between 1 and 12 s, while 0.5–3 s can achieve better classification accuracy (Li et al., 2017). For the DEAP dataset, each subject has 40 × 60 s of emotional EEG signals, and we set the size of the sliding window to 2 s without overlapping, so that a total

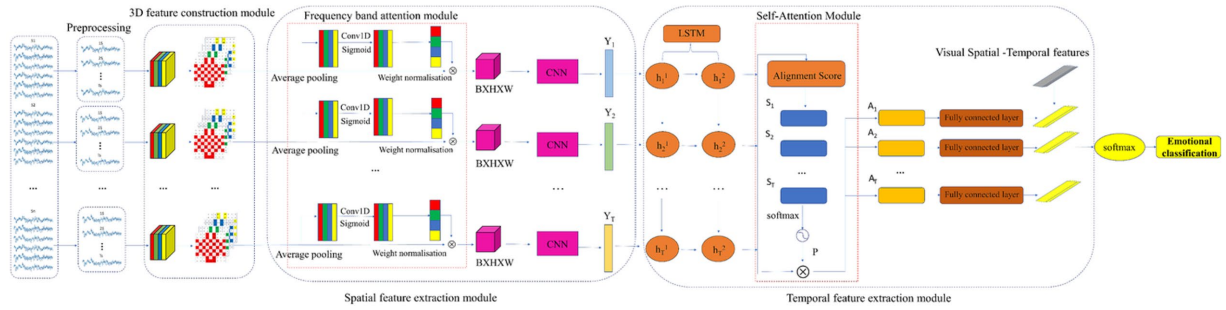

FIGURE 3
Framework for the EEG model.

of 26,400 samples can be obtained, with 1,200 samples for each person; while for the MAHNOB-HCI dataset, due to the varying durations of each trial, we choose the middle 30 s of each video as the experimental video data, and set the size of the sliding window to 0.5 s without overlapping, so that a total of 30,000 samples can be obtained, with 1,200 samples for each person.

The pre-processed EEG signals were fed into the 3D feature construction module. Each time-slice sample was first decomposed into four frequency bands, i.e., θ (4–7 Hz), α (8–13 Hz), β (14–30 Hz) and γ (31–50 Hz), using a Butterworth filter (Zheng and Lu, 2015), and then the differential entropy (DE) features of these four bands were calculated separately. The researchers found that the differential entropy feature is currently the most effective feature in the field of emotion recognition (Chen et al., 2019), and the formula is shown in Eq. (1).

$$D(x) = -\int_{x} f(x) log f(x) dx \tag{1}$$

where $f(x)$ is the probability density function of x. According to Zheng et al. the differential entropy characteristic formula for the Gaussian distribution is shown in Eq. (2), where e is Euler's constant and σ is the standard deviation of the EEG sequence.

$$h(z) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(z-u)^2}{2\sigma^2} \log\frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(z-u)^2}{2\sigma^2} dz = \frac{1}{2}\log 2\pi e\sigma \tag{2}$$

The differential entropy features of each frequency band were then projected onto a two-dimensional matrix (Li et al., 2018; Nguyen et al., 2019; Sha et al., 2023), with the length and width of the two-dimensional matrix set to H = 9 and W = 9, respectively, and the relative positions of the actual recording electrodes corresponding to the positions of the recording electrodes in the two-dimensional matrix. Figure 4 shows the two-dimensional matrix obtained from the projection based on 32 sampled electrodes, with the unused channel signals filled with zeros. Finally, the four frequency bands of each EEG sample were stacked to obtain a three-dimensional feature representation of each EEG signal and as shown in Eq. (3):

$$E_i = [E_1, E_2, E_3 \ldots E_T] \in R^{T \times B \times H \times W} \tag{3}$$

2.2.2 Frequency band attention module

The band attention module used in this paper is inspired by the ECAnet Convolutional Attention Module (Han et al., 2021) in the field of image recognition, and uses a one-dimensional convolution to interact with the information in each band, with the size of the convolution kernel varied by an adaptive function. Specifically, for the EEG sample $E_i \in R^{B \times 9 \times 9}$, the matrix with feature maps [B,H,W] is first converted to a vector of [1,c] by a global average pooling layer, and then the one-dimensional convolution kernel size kernel_size is obtained by an adaptive function, the formula of which is shown in Eq. (4):

$$k = \left| \frac{\log(c)}{y} + \frac{b}{y} \right| \tag{4}$$

Where $y=2$ and $b=1$, We calculate the size of kernel_size and apply it to the one-dimensional convolution, then multiply to [1,c] reshape into [c,1] and multiply by with the one-dimensional convolution to get the weight for each band in the feature map, and finally normalize the weights and multiply with by the original feature map to get the weighted feature map and as shown in Eq. (5):

$$E_i^* = E_i \otimes Sigmoid\left(\mathrm{cov}1D(filters = 1, Kenel\_size = c)\right)(x) \tag{5}$$

2.2.3 The CNN-LSTM networks

The CNN-LSTM networks consists of four successive convolutional layers, a maximum pooling layer, a fully connected layer and an LSTM layer. There are four successive convolutional layers with convolutional kernel sizes of 5×5, 4×4, 3×3 and 1×1, and output channels of 64, 128, 256 and 128, in respectively, all of which apply the zero-filling and RELU activation functions are applied. A maximum pooling layer with convolutional kernel size 2×2 and step size 2 is used to improve the robustness of the network, the output of the pooling layer is flattened and fed into the fully connected layer, which outputs 512 units, and Et is set as the input of the CNN, $E_t \in R^{1 \times B \times H \times W}$, and Eqs. (6)–(11) are used to describe the computation of the layers in the CNN:

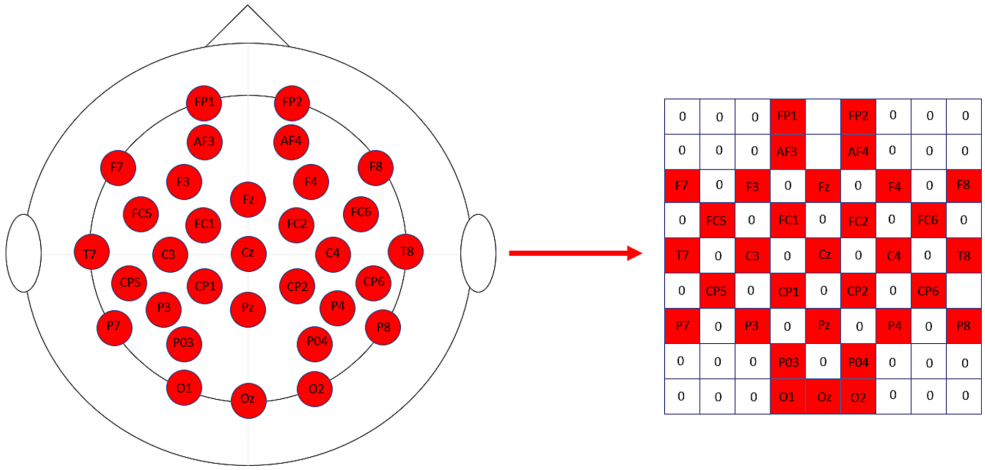

FIGURE 4
The 32 sampled electrodes are projected onto the 2D map on the right according to their relative positions.

$$C_1 = f\left(Conv\left(E_t, w_{c_1}\right)\right), W_{c_1} \in R^{5\times5} \tag{6}$$

$$C_2 = f\left(Conv\left(C_1, w_{c_2}\right)\right), W_{c_2} \in R^{4\times4} \tag{7}$$

$$C_3 = f\left(Conv\left(C_2, w_{c_3}\right)\right), W_{c_3} \in R^{3\times3} \tag{8}$$

$$C_4 = f\left(Conv\left(C_3, w_{c_4}\right)\right), W_{c_4} \in R^{1\times1} \tag{9}$$

$$C_5 = f\left(Conv\left(C_4, w_{c_5}\right)\right), W_{c_5} \in R^{2\times2} \tag{10}$$

$$y_t = linea\left(flatten\left(C_5\right)\right) \tag{11}$$

Where $f(\cdot)$ denotes the Relu activation function, $W_{c1}, W_{c2}, W_{c3}, W_{c4}, W_{c5}$ denote the convolution kernel of each convolutional layer, and E_t is the input matrix, and we input E_t into the spatial convolutional network to obtain the spatial feature representation of E_t, y_t. where, y_t t = 1,2,…T denotes the feature vector from the 1st sample, 2nd sample to the Tth sample.y* is obtained by concatenating all the feature vectors y_t in chronological order.

The CNN output sequence is $y^* = (y_1, y_2, y_3 \ldots y_t$, where $y_t \in R^{1\times512}$, t = 1,2,3. y_t, where $y_t \in R^{1\times512}$, t = 1,2,3… T), y* is input to the LSTM layer, the number of LSTM layers is set to 2, and the number of hidden units is set to the sample number, so it can be considered that the output of each time step is the spatial–temporal feature information of each sample, and the output of the LSTM network is computed as shown in Eqs. (12)–(16):

$$i_t = \sigma\left(W_{yi}y_t + W_{hi}h_{t-1} + W_{ci}c_{t-1+}b_i\right) \tag{12}$$

$$f_t = \sigma\left(W_{yf}y_t + W_{hf}h_{t-1} + W_{cf}c_{t-1+}b_f\right) \tag{13}$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{yc}y_t + W_{hc}h_{t-1} + b_c\right) \tag{14}$$

$$o_t = \sigma\left(W_{yo}y_t + W_{ho}h_{t-1} + W_{co}c_{t+}b_o\right) \tag{15}$$

$$h_t = o_t \tanh\left(c_t\right) \tag{16}$$

where σ is the logical sigmoid activation function, i, f, and o are the input, forgetting, and output gates, respectively, and C is the cell activation vector. ht. denotes the Tth output hidden state of the seconds recursive layer and its expression is shown in Eq. (17):

$$h_t \mid h_t = Lstm\left(y_t\right), t = 1,2,3\ldots T, h_t \in R^{T\times128} \tag{17}$$

2.2.4 Self-attention module

As shown on the right side of Figure 3, the self-attention module aims to assign different weights to each EEG sample in order to explore the importance between different samples and to extract more discriminative spatial–temporal feature information. The feature score vector S_t is first computed for each hidden state h_t and the formula is shown in Eq. (18):

$$S_t = f\left(h_t, d_t\right) = W_t\sigma\left(W_1h_t + W_2d_t + b_1\right) + b_2 \tag{18}$$

where $f(t)$ denotes the importance of the Tth coded sample, and d_t is the aligned pattern vector generated from h_t by linear transformation with the same dimension as h_t. The activation function is set to Relu, W_t and b_2 denote the weight matrix and bias term of the activation function, respectively. W_1, W_2 are the weight matrices of h_t and d_t and b_1 is the bias term. The similarity N_t within each sample of different points is obtained by dot-multiplying the transpose S_t^T of the feature score vector with the output hidden state h_t The probabilistic representation of N_t by SoftMax function, the probability of the Tth hidden layer state can be expressed as shown in Eqs. (19), (20):

$$N_t = s_t^T h_t \tag{19}$$

$$P_t = \frac{\exp(N_t)}{\Sigma_{t=1}^{T} \exp(N_t)} \tag{20}$$

Each output hidden layer h_t is then allowed to multiply with its computed probability to obtain the features extracted by the self-attention module: $A = \{A_1, A_2, A_3 \ldots A_t\}$, $t = 1,2,3 \ldots T$. Finally, the extracted spatial–temporal attention features are fed into a classifier consisting of a fully connected layer and a SoftMax layer to output the final emotion type labels.

2.3 EEG network enhanced by visual network

We use the knowledge gained from the visual network to improve the performance of the EEG network. Firstly, the facial expression video frames in the emotion database are used to pre-train the visual model, then, the trained visual network is used to extract the features of the visual modality in the target dataset, and finally, the raw EEG signals together with the corresponding labels and features obtained from the visual network are input into the EEG network in an offline manner, and the EEG network is trained using the weighted sum of the cross-entropy function and the L_1 loss function as the loss function to make the whole The training process is more controllable, and its formula is as shown in Eq. (21):

$$L_c = -\rho \sum_{t=1}^{T} Y_t \log(P_t) + (1-\rho) L_1(V_t - V_s) \tag{21}$$

where Y_t denotes the label of the Tth EEG sample, P_t denotes the predicted probability of the Tth sample, V_t and V_s denote the spatial–temporal features of the visual and EEG networks, respectively, and ρ is a hyperparameter that is manually set to 0.8. The cross-entropy loss function is the primary loss function, which updates the weight matrix w in the model by the discrepancy between the actual prediction and the expected label, and reduces the distance between the actual prediction and the expected label, and a lower cross-entropy loss function represents a higher emotion recognition accuracy; the L_1 loss function is the auxiliary loss function, inspired by Romero et al. (2015), which extracts knowledge by enforcing the proximity of the intermediate features graphs using the L_1 loss function, and the L_1 loss function is computed as shown in Eq. (22):

$$L_1 = w * \frac{1}{TF} \sum_{t=1}^{T} |U_i - V_i| \tag{22}$$

where $U_i \in R^{T \times F}$ and $V_i \in R^{T \times F}$ denote the feature sequences obtained at each time step, and w^* is the hyperparameter, and the optimal w^* is found by grid search. The multi-task loss function L_c guides the training process of the EEG network so that the EEG network can learn the Knowledge of the visual network, thus achieving the purpose of improving the emotion recognition performance of the EEG network using the visual modality.

3 Results

3.1 Introduction to source datasets

To evaluate the efficacy of the proposed network model, we conducted experiments on two multimodal data sets, DEAP and MAHNOB-HCI; Table 1 provides the relevant details for the two datasets.

The MAHNOB-HCI is a multimodal emotion database of 30 young healthy adult participants, 17 females and 13 males. Age ranging from 19 to 40 years (M = 26.06 SD = 4.39). The MAHNOB-HCI was used to record responses to emotional stimuli and to record facial expression videos, audio signals, EEG signals and other physiological signals from the 30 participants. After viewing 20 emotional video clips, participants rated their emotional experience on each of the four dimensions of arousal, valence, control and predictability, labeling the dimensions of arousal, valence, control and predictability on a scale of 1–9 on each trial. Facial expression videos were transmitted at 60 frames per second. Due to problems with the experimental equipment or the experimental recording, data were incomplete for five individuals, so the actual number of participants in our experiment was 25, with 20 trials per person for each dimension.

The DEAP dataset is a multimodal dataset for the analysis of human emotional states. EEG and peripheral physiological signals were recorded from 32 participants. 22 of the 32 participants recorded frontal videos using a Sony DCR-HC27E camcorder, so we used the data from these 22 subjects. Subjects watched 40 one-minute music video clips and rated their emotional experience on five dimensions: arousal, valence, dominance, liking and familiarity, on a discrete scale of 1–9, except for familiarity, which was rated on a discrete scale of 1–5. Setting the transmission speed of facial expression videos was set from the original 50 to 60 fps, which is the same as the MAHNOB-HCI dataset, to facilitate subsequent unified processing.

We conducted subject-related emotion recognition experiments on the DEAP dataset and the MAHNOB-HCI dataset to assess the feasibility of the proposed method.

Initially, we examined the efficacy of the 3FACRNN network in identifying emotions using the DEAP dataset and the MAHNOB-HCI dataset. The emotion categories for each trial were hierarchically designated along the dimensions of arousal and valence, respectively, using the subjects' own levels of arousal and valence from the DEAP

TABLE 1 Detailed information on the DEAP dataset and the MAHNOB-HCI dataset.

Item	DEAP	MAHNOB-HCI
Subjects	22	25
Trail of each subject	40	20
Each clip duration	60s	30s
Available channels	32	32
Sampling rate	128 Hz	256 Hz
Items for rating emotion	Valence, Arousal	Valence, Arousal

and MAHNOB-HCI datasets as the criteria for self-rating emotions. On a scale from 1 to 9, participants rated their arousal and valence, and we chose 5 as the threshold to divide the labels into two binary classification problems. The overall performance of the approach was evaluated by considering the average classification accuracy, precision, recall, and F1 scores across all participants. Next, we performed ablation experiments on the 3FACRNN network to examine the impact of the 3D construction module, attention module, and visual modality on the classification accuracy. Additionally, we calculated the attentional weights of the various bands to assess the significance of each band in the emotion recognition process. Ultimately, we compared the 3FACRNN network model with previously reported methods for the DEAP dataset and the MAHNOB-HCI dataset.

The models used in this paper are implemented by Openface, Keras2.6.0, Keras2.6.0 is extended by Tensorflow2, all model training is performed on NVIDIA GeForce RTX 3060 laptop GPU.

3.2 Emotion recognition using 3FACRNN

In order to train the 3FACRNN network model, we initially trained the visual network on the AFEW-VA dataset. This was done as a fine-tuning step for the facial expression recognition task. The learning rate for the visual model was set to 1e-5, the maximum number of epochs was set to 100, and the batch size was set to 128. Additionally, the optimal model parameters were loaded at the end of each epoch. The learning rate for the EEG network model was set to $1e-6$. The maximum number of epochs was set to 100, and the batch size was set to 128. A grid search was performed using Eq. (22), with the parameter w ranging from 0.5 to 1.5 and a step size of 0.1. The hyper-parameters were optimized using the test set. Five times tenfold cross-validation is applied on each subject. The average classification accuracy and standard deviation of all subjects were used as the final results to represent the performance of the 3FACRNN network model.

The 3FACRNN network produced the greatest recognition results when the grid search parameter $w = 1.0$. The average recognition accuracy and standard deviation of the 3FACRNN network for all subjects in the DEAP dataset were 96.75 ± 1.75 and 96.86 ± 1.33. This is marginally inferior to the performance on the MAHNOB-HCI dataset, and it is possible that this is due to the fact that the mood induction situation varied between subjects. In addition, for a comprehensive evaluation of the performance of the 3FACRNN network, the F1 Score, a reconciled average of precision and recall, is used as the network model evaluation result. Table 2 demonstrates that the F1 scores of the 3FACRNN network on the emotion and arousal dimensions of the two datasets are 96.09, 96.34, 97.38, and 97.33, respectively. All F1 Scores are greater than 96%, indicating that the 3FACRNN network achieves satisfactory classification results on both datasets. Figure 5 present the accuracies for all subjects in the DEAP dataset and the MAHNOB-HCI dataset, revealing that the 3FACRNN network can achieve more accurate classification results for all subjects in the two datasets, thereby demonstrating its superiority on the two datasets. Figure 6 depict the confusion matrices derived for the 3FACRNN network on the DEAP dataset and the MAHNOB-HCI dataset. Figure 6 demonstrate that the recognition rate of the 3FACRNN network is greater for low valence and low arousal samples than for high valence and high arousal samples, and that the 3FACRNN network can achieve the optimal classification of the emotions for low valence and low arousal samples.

TABLE 2 Accuracy, precision, recall and F1 score obtained by 3FACRNN network on DEAP dataset and MAHNOB-HCI dataset.

Dataset	DEAP		MAHNOB-HCI	
Index	Arousal	Valence	Arousal	Valence
Accuracy	96.75%	96.86%	97.55%	98.37%
Precision	97.57%	97.95%	98.89%	99.13%
Recall	94.66%	94.79%	96.42%	97.26%
F1 score	96.09%	96.34%	97.38%	97.33%

3.3 Ablation experiment

In this paper, we validate the effects of multiple attentional mechanisms in the 3FACRNN network model on an emotion recognition task and design four models to compare their performance in an emotion EEG recognition task: the first model contains both the banded attention module and the self-attention module, the second model contains only the banded attention module, the third model contains only the self-attention module, and the fourth model does not contain any attention module. The purpose of constructing these four models was to verify the validity of the self-attention module and the frequency band attention module. Table 3 displays the mean accuracy and standard deviation for each of the four models.

To investigate the effect of various attention modules on the performance of the 3FACRNN network, we compare and analyze the mean accuracy of each network model in Table 3. The first network model, which includes all attention modules, has the greatest improvement in recognition accuracy compared to the fourth network model on the arousal and valence dimensions of the DEAP and MAHNOB-HCI datasets, with improvements of 6.64%, 7.43%, 9.02%, and 7.19%, respectively. With the addition of the self-attention module, the average accuracy of the first network model increased by 1.91%, 2.51%, 3.33%, and 4.22% when compared to the third model. The recognition performance of the network model with the addition of the self-attention module alone is superior to that of the network model with the addition of the frequency-band attention module alone. This is because the frequency-band attention module captures the feature information of different frequency bands in the EEG samples from the local time-slice domain, while The self-attention module captures the intrinsic attentional information between samples.

In order to determine the significance of the 3D feature construction module in the 3FACRNN network model, we input the raw EEG signals directly into the spatial–temporal convolutional network to extract the local spatial–temporal features, bypassing the 3D feature construction module. Because the differential entropy characteristics of the frequency bands in the EEG signal are not utilized in this procedure, the frequency band attention module in the 3FACRNN network is also eliminated, while the other inherent network structures are preserved. First, the raw EEG signals are maintained, then each sampling point is projected onto a spatial matrix, and lastly, the signal matrix is fed directly into a spatial–temporal convolutional network for feature extraction and

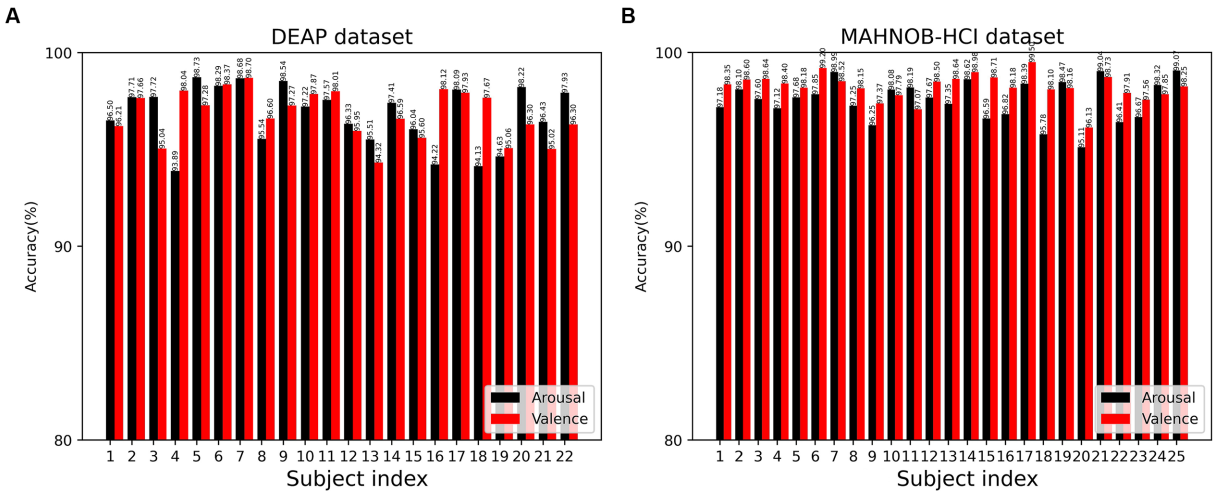

FIGURE 5
Recognition accuracy of the 3FACRNN network on the DEAP dataset and MAHNOB-HCI dataset for all subjects: **(A)** DEAP dataset; **(B)** MAHOB-HCI dataset.

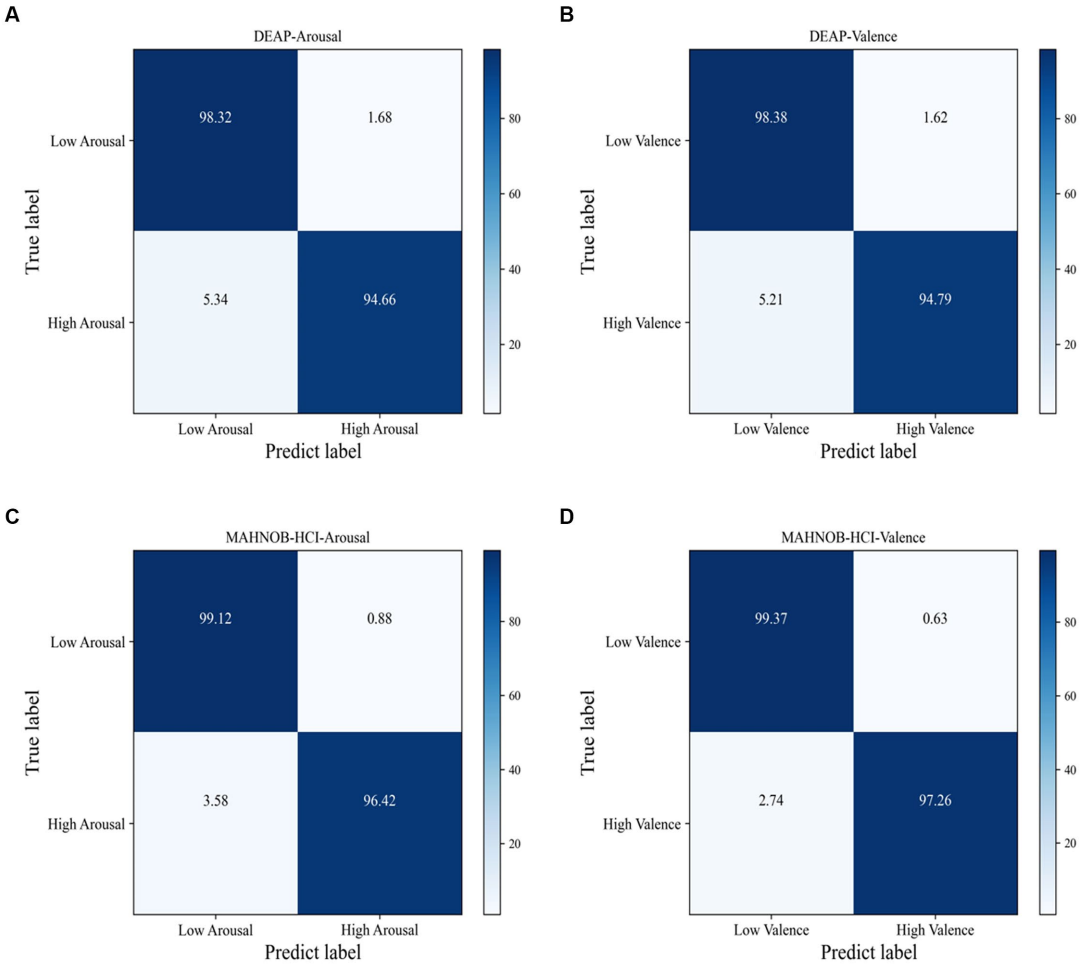

FIGURE 6
Confusion matrix of the proposed 3FACRNN on DEAP dataset and MAHNOBHCI dataset: **(A)** DEAP:Arousal; **(B)** DEAP:Valence; **(C)** MAHNOBHCI:Arousal; **(D)** MAHNOBHCI:Valence.

TABLE 3 Average accuracy and standard deviation obtained by the 3FACRNN network for different attention situations.

Attention	DEAP		MAHNOB-HCI	
	Arousal	Valence	Arousal	Valence
With all attention	96.75 ± 1.75%	96.86 ± 1.33%	97.55 ± 1.51%	98.37 ± 1.07%
With only Frequency Band-attention	92.61 ± 3.91%	93.59 ± 3.55%	93.61 ± 3.91%	93.59 ± 3.65%
With only Self-attention	94.84 ± 2.61%	94.35 ± 2.89%	94.22 ± 0.2.37%	94.15 ± 2.46%
W/O any attention	90.11 ± 3.58%	89.43 ± 4.49%	88.53 ± 3.74%	91.18 ± 4.66%

TABLE 4 Comparison of recognition performance of 3FACRNN networks with raw EEG signal as input and processed by 3D constructor module as input.

Input EEG signals	DEAP		MAHNOB-HCI	
	Arousal	Valence	Arousal	Valence
With 3D-feature structure	96.75 ± 1.75%	96.86 ± 1.33%	97.55 ± 1.51%	98.37 ± 1.07%
With raw EEG signals	94.22 ± 3.12%	93.80 ± 2.69%	93.99 ± 2.57%	93.74 ± 2.32%

TABLE 5 Comparison of recognition performance of 3FACRNN networks with and without visual pattern involvement.

Visual feature	DEAP		MAHNOB-HCI	
	Arousal	Valence	Arousal	Valence
With visual feature	96.75 ± 1.75%	96.86 ± 1.33%	97.55 ± 1.51%	98.37 ± 1.07%
W/O visual feature	93.47 ± 3.01%	93.18 ± 3.83%	92.46 ± 3.36%	94.54 ± 3.02%

classification. The average accuracy and standard deviation of the two datasets for the valence and arousal dimensions are displayed in Table 4. Table 4 reveals that the average accuracy and standard deviation of the network model on the two datasets are 94.22 ± 3.12, 93.80 ± 2.69, 93.99 ± 2.57, and 93.74 ± 2.32, with the raw EEG signals as inputs, and that it is lower than the average accuracy of the 3FACRNN network model by 2.53%, 3.06%, 3.56%, and 4.63%, respectively. Experiments have shown that adding the 3D feature construction module to the 3FACRNN network can improve recognition accuracy. This is because the EEG signals processed by the 3D feature construction module are more complex at the feature level and contain more useful emotional information.

To investigate the effect of visual modalities on the recognition accuracy of the network, we conducted experiments without including visual modalities. Table 5 shows the average accuracy and standard deviation of the network without and with visual modalities. In Table 5, the average accuracy with standard deviation of the network on the two datasets without considering visual modality is 93.47 ± 3.01, 93.18 ± 3.83, 92.46 ± 3.36, and 94.54 ± 3.02, and it is lower than that of the 3FACRNN network considering visual modality by 3.28%, 3.68%, 5.09%, and 3.83%, respectively. This due to the fact that multimodality captures more comprehensive feature information in the global time domain than unimodality, and experimental results show that allowing the EEG modality to learn the dark knowledge of the visual modality improves the recognition performance of the 3FACRNN network.

3.4 Analysis of the attention weighting of the frequency average band

To comprehend the significance of different frequency bands in the emotion recognition process, we calculated the average frequency band attentional weight values of all subjects after training, which represents the significance of different frequency bands in the network training process, and plotted the average frequency band weight rectangles of the four frequency bands in Figure 7. Because the attention weights for each frequency band have been normalized, Figure 7 displays values in the range [0,1] for the attention weights for the four frequency bands. In the DEAP and MAHNOB-HCI datasets, the network assigned the highest attentional weights to the gamma band. Since the network continuously updates the band attentional weights during training, this also suggests that the differential entropy feature of the gamma band provides a more discriminative feature during emotion recognition. Due to the shorter EEG sample duration of subjects in the MAHNOB-HCI dataset, the average band attentional weights of the four bands are lower in the MAHNOB-HCI dataset than in the DEAP dataset.

3.5 Method comparison

We compared the proposed 3FACRNN network model with the state-of-the-art methods on the DEAP dataset and MAHNOB-HCI dataset, as shown in Table 6, with a brief description of each method as follows:

1 DBN (Wang and Shang, 2013): A Deep Belief Network (DBN)-based emotion recognition system that automatically extracts features from four channels of raw EEG data in an unsupervised manner and accomplishes emotion classification.
2 M-CLASS (Sander and Ioannis, 2013): A Multimodal Emotion Recognition Method for Emotion Recognition after Fusion of Facial Expression Features and EEG Features at Decision Layer or Feature Layer.

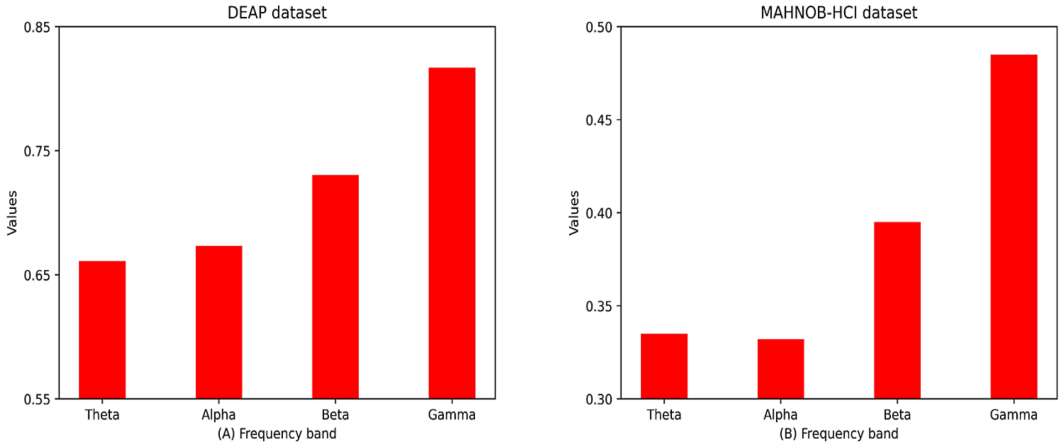

FIGURE 7
Mean spectral attention masks for all subjects in the DEAP and MAHNOB-HCI datasets in four frequency bands (i.e., theta, alpha, beta, and gamma frequency bands):**(A)** DEAP dataset; **(B)** MAHOB-HCI dataset.

TABLE 6 Comparison of mean accuracy and standard deviation (acc ± std.%) between the baseline method and the proposed 3FACRNN network on the DEAP dataset and the MAHNOB-HCI dataset.

Author	Methods	Year	DEAP		MAHNOB-HCI	
			Arousal	Valence	Arousal	Valence
D. Wang	DBN	2013	60.9	51.2	-	-
K. Sander	M-CLASS	2013	-	-	66.5	71.5
T. F. Song	GCNN	2018	87.72 ± 3.32	88.24 ± 3.18	-	-
Y. Yang	Conti-CNN	2018	81.55 ± 6.55	82.77 ± 4.47	-	-
D. Zhang	CRAM	2019	84.46 ± 9.27	87.09 ± 7.49	-	-
Y. G. Huang	Multi-CNN	2019	-	-	74.17	75.21
J. Chen	CNN-LSTM	2020	93.26	93.64	-	-
F. Shen	4D-CRNN	2020	94.58 ± 3.69	94.22 ± 2.61	-	-
J. Liu	CSDNN	2020	92.86	89.49	-	-
Z. Wang	MDBN	2020	87.32	83.69	-	-
X. L. Zhong	MA-attention	2020	-	-	70.25	73.27
Z. Gao	DCNN	2021	92.92	92.24	-	-
X. Deng	SFENet	2021	91.94	92.49	-	-
Y. Yin	GCN-LSTM	2021	90.60	90.45	-	-
Siddharth	Deep learning	2022	-	-	80.42	80.77
Yong Zhang	HC-MFB	2022	-	-	90.37	90.50
C. Li	CADD-DCCNN	2023	92.42	90.97	-	-
G. Q. Peng	TR&CA	2023	95.58 ± 2.28	95.18 ± 2.46	-	-
Ours proposed	3FACRNN	2023	96.75 ± 1.75	96.86 ± 1.33	97.55 ± 1.51	98.37 ± 1.07

3 Conti-CNN (Yang et al., 2018b): A three-dimensional input continuous convolutional neural network combining features from multiple bands to improve the accuracy of emotional EEG recognition.
4 CRAM (Zhang et al., 2019): An emotion recognition network that uses CNNs to abstractly encode EEGs and a recursive attention mechanism to extract spatial–temporal features in EEGs for emotion classification.
5 GCNN (Song et al., 2020): A network that uses spectrogram filtering to extract different differential entropy features for emotional EEG recognition.
6 CNN-LSTM (Chen et al., 2020): An emotional EEG signal recognition network using a hybrid convolutional recursive module of CNN and LSTM.
7 4D-CRNN (Shen et al., 2020): A four-dimensional convolutional recurrent neural network is proposed to convert

the differential entropy features of different channels into a 3D structure to train the network model.

8 CSDNN (Liu J. et al., 2020): An emotion recognition network combining convolutional neural networks with sparse autoencoders.

9 MDBN (Wang et al., 2020): Multimodal emotion recognition using deep belief networks.

10 DCNN (Gao et al., 2021): A dense convolutional neural network for sentiment recognition using channel fusion methods.

11 SFENet (Deng et al., 2021): An emotion recognition network based on spatial folding integration.

12 GCN-LSTM (Yin et al., 2021): An emotion recognition algorithm based on graph convolutional neural networks and long and short-term memory neural networks.

13 Multi-CNN (Huang et al., 2019):A deep convolutional neural network combining facial expressions and EEG for enhanced emotion recognition.

14 MA-attention (Zhong et al., 2020): A convolutional neural network using moving average (MA) and attentional mechanisms was designed to recognize emotional EEG signals.

15 Deep learning (Siddharth and Jung, 2022): A deep learning method is designed to implement a multimodal vision and EEG based affective computing network using deep learning methods.

16 HC-MFB (Zhang et al., 2022):A multimodal emotion learning network model based on heterogeneous convolutional neural networks and multimodal factorized bilinear pools is designed.

17 CADD-DCCNN (Li et al., 2023): A causal convolutional neural network based on cross-attention mechanism for EEG emotion recognition.

18 TR&CA (Peng et al., 2023): An emotion recognition network based on channel attention mechanism and time relative coding mechanism.

Table 6 reports the comparison of the recognition performance of all the above methods and the proposed 3FACRNN network in this paper on DEAP and MAHNOB-HCI datasets. Overall, the proposed 3FACRNN network outperforms the state-of-the-art methods with average recognition accuracies of 96.75, 96.86, 97.55, and 98.37 on DEAP and MAHNOB-HCI datasets, respectively, with standard deviations of 1.75, 1.33, 1.51, and 1.07, respectively.

We compared the 3FACRNN network with the Conti-CNN method proposed by Yang et al. The 3FACRNN network outperforms the Conti-CNN method in average recognition accuracy by 15.2% and 14.09%, respectively, which is due to the fact that Yang et al. only considered the feature information of the EEG signals in terms of spatial domain, and did not take into account the EEG signals in terms of time domain feature cues, while the 3FACRNN network utilizes the LSTM network to obtain the long-term temporal features of emotional EEG signals, which improves the network's prediction performance of emotional states.

Although Chen et al. used a hybrid CNN and LSTM convolutional recurrent neural network to extract the feature information of EEG signals in both spatial and temporal domains, the average recognition accuracy was still 3.49% and 3.22% lower than that of the 3FACRNN network, which is due to the fact that the 3FACRNN network not only uses a convolutional recurrent neural network based on CNN and LSTM to extract the spatial and temporal feature information of the EEG signals but also incorporates a frequency band attention module and a self-attention module to enhance the discriminative property of the feature information. In addition, Zhang et al. incorporated a recursive attention mechanism into a convolutional neural network to explore the effect of different time-slice samples on the emotion recognition process, but their average recognition accuracy was lower than that of the method proposed in this paper, which further illustrates the effectiveness and advanced nature of the multi-attention mechanism chosen in this paper.

Yin et al. used EEG signals obtained from different electrode channels to construct a brain network, and adopted the brain network representation learning method of graph neural network to obtain the feature representation of EEG signals in spatial and temporal dimensions, and finally extracted the temporal features of emotional responses using LSTM network, which can achieve an average recognition accuracy of 90.60% and 90.45%, but since graph neural network needs to perform the feature vector computation while adjusting the structure between brain network graphs, so the efficiency of the algorithm will show a significant decrease with the increase of feature graphs. The 3FACRNN network spatially projected the EEG features in order to maintain the relative relationship between the placement of EEG electrodes on the head. The EEG features of each frequency band were first mapped into a 2D matrix and then organized into a 3D structure according to the frequency bands, the 3FACRNN network outperformed the GCN-LSTM method of Yin et al. by 6.15 and 6.41%, respectively, in terms of average recognition accuracy, and also outperformed the GCN-LSTM method in terms of algorithmic efficiency.

Siddharth et al. used deep convolutional network to extract the emotional feature information of visual modality and EEG modality and combined the two feature information for the prediction of emotional state with an average recognition accuracy of 74.17% and 75.21%. Zhang et al. proposed a multimodal emotion learning network model based on heterogeneous convolutional neural network and multimodal factorized bilinear pool. The network model fuses the feature information of visual modality and EEG modality in the decision layer, and the average recognition accuracy can reach 90.37% and 90.50%. In this paper, we propose a multimodal emotion recognition network, 3FACRNN, which utilizes a multitask loss function Lc to force approximation of intermediate feature vectors of visual and EEG modalities in order to improve the recognition performance of the EEG network through visual knowledge. The 3FACRNN network takes into account the differences between different modal features, and its average recognition accuracy is higher than that of the Zhang et al. and Siddharth et al. proposed methods.

4 Discussion

In this paper, we conducted extensive experiments using 3FACRNN networks on two public datasets and obtained satisfactory performance, and we next discuss points in the 3FACRNN networks that can be further refined.

We add a 3D feature construction module to the 3FACRNN network, which projects the electrode position information of the EEG samples into a 2D matrix to facilitate the subsequent convolution operation. We set the length and width of the 2D matrix to 9×9. We designed a more compact 2D matrix compared to the sparse matrix used by Li et al. The compact matrix has a smaller size and requires relatively fewer convolutional kernels, consuming less time cost, while the sparse matrix requires more convolutional kernels to extract more features from the EEG samples, which is more favorable for the subsequent recognition classification task. So next we will investigate the application of sparse maps in the field of multimodal EEG sentiment recognition.

From Table 6, we can see that the 3FACRNN network outperforms the CRAM method in both dimensions of the DEAP dataset, producing a significant increase of 12%, which can be attributed to the fact that the convolutional recurrent neural network of the 3FACRNN network is much deeper, containing four convolutional layers, one pooling layer, one linear layer and one LSTM layer, and producing a feature map with {64,128,256,128} feature maps, whereas the CRAM method contains only one convolutional layer, one pooling layer and one LSTM layer, producing {40} feature maps, and its network depth is much lower than that of the 3FACRNN network. Deeper convolutional and pooling layers also allow the 3FACRNN network to extract and retain more emotion-related cues.

5 Conclusion

In this paper, we propose a 3FACRNN network model for multimodal emotion recognition, which includes two parts, the EEG network and the visual network. A 3D feature construction module was added to the EEG network with the aim of complexly extracting the electrode information, frequency band information and spatial–temporal information from the original EEG signal to provide more feature cues to the convolutional recurrent network; In addition, we used the frequency band attention module and the self-attention module to make the feature information extracted by the convolutional recurrent network more discriminative at both local and global time slice scales. Finally, the two network models reduce the proximity of the intermediate feature maps through a multi-task loss function L_C, which allows the EEG network to learn the knowledge of the already trained visual network and improves the performance of the EEG network for affective computing. We pre-train the visual models using CNN and TCN, and then use the spatial–temporal features in the trained visual networks as dark knowledge to improve the recognition performance of EEG networks. The experimental results demonstrate the effectiveness of our proposed 3FACRNN network model. The 3FACRNN network can understand the feature information of different modalities, and it can also complexify the feature information through the 3D feature construction module and the multi-attention mechanism, so as to make it contain more information conducive to the recognition of emotions, and to improve the recognition performance of the network. In future work, we will apply the 3FACRNN network to topic-independent and cross-session tasks to improve the generalization of the model.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

YD: Data curation, Methodology, Software, Writing – original draft, Conceptualization, Investigation, Visualization. PL: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Writing – review & editing. LC: Investigation, Methodology, Validation, Writing – review & editing, Conceptualization. XZ: Supervision, Writing – review & editing, Formal Analysis. ML: Investigation, Validation, Writing – review & editing. FL: Investigation, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (nos. 62271350 and 61806146), the National Key Research and Development Program of China (grant no. 2021YFF1200600), the Natural Science Foundation of Tianjin City (nos. 18JCYBJC95400 and 19JCTPJC56000).

Acknowledgments

The authors sincerely thank the editors and reviewers for their constructive suggestions.

Conflict of interest

LC was employed by China Electronics Cloud Brain (Tianjin) Technology Co, Ltd., Tianjin.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Afouras, T., Chung, J. S., and Zisserman, A. (2020), ASR is all you need: cross-modal distillation for lip Reading, In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2143–2147

Ahmed, M. Z. I., Sinha, N., Ghaderpour, E., Phadikar, S., and Ghosh, R. (2023). A novel baseline removal paradigm for subject-independent features in emotion classification using EEG. *Bioengineering*. 10:21. doi: 10.3390/bioengineering10010054

Basbrain, A., and Gan, J. Q. (2020), One-shot only real-time video classification: a case study in facial emotion recognition. 197–208

Blankertz, B., Acqualagna, L., Dähne, S., Haufe, S., Schultze-Kraft, M., Sturm, I., et al. (2016). The Berlin brain-computer Interface: Progress beyond communication and control. *Front Neuro Sci* 10:530. doi: 10.3389/fnins.2016.00530

Chen, J., Jiang, D., Zhang, Y., and Zhang, P. (2020). Emotion recognition from spatialtemporal EEG representations with hybrid convolutional recurrent neural networks via wearable multi-channel headset. *Comput. Commun.* 154, 58–65. doi: 10.1016/j.comcom.2020.02.051

Chen, D. W., Miao, R., Yang, W. Q., Liang, Y., Chen, H. H., Huang, L., et al. (2019). A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition. *Sensors* 19:17. doi: 10.3390/s19071631

Cheng, J., Chen, M. Y., Li, C., Liu, Y., Song, R. C., Liu, A. P., et al. (2021). Emotion recognition from Multi-Channel EEG via deep Forest. *IEEE J. Biomed. Health Inform.* 25, 453–464. doi: 10.1109/jbhi.2020.2995767

D'mello, S., and Kory, J. (2015). A review and Meta-analysis of multimodal affect detection systems. *ACM Comput Surv.* 47, 1–36. doi: 10.1145/2682899

Daros, A. R., Zakzanis, K. K., and Ruocco, A. C. (2013). Facial emotion recognition in borderline personality disorder. *Psychol. Med.* 43, 1953–1963. doi: 10.1017/s0033291712002607

Deng, X., Zhu, J., and Yang, S. (2021), SFE-net: EEG-based emotion recognition with symmetrical spatial feature extraction, In: *Proceedings of the 29th ACM international conference on multimedia, Association for Computing Machinery*, Virtual Event, China. pp. 2391–2400

Doma, V., and Pirouz, M. (2020). A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals. *J. Big Data*. 7:21. doi: 10.1186/s40537-020-00289-7

Gao, Z., Wang, X., Yang, Y., Li, Y., Ma, K., and Chen, G. (2021). A channel-fused dense convolutional network for EEG-based emotion recognition, IEEE trans. *Cogn. Dev. Syst.* 13, 945–954. doi: 10.1109/TCDS.2020.2976112

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., et al. (2022). Attention mechanisms in computer vision: A survey. arXiv [Preprint] 8. 331–368

Han, G., Zhang, M., Wu, W., He, M., Liu, K., Qin, L., et al. (2021). Improved U-net based insulator image segmentation method based on attention mechanism. *Energy Rep.* 7, 210–217. doi: 10.1016/j.egyr.2021.10.037

He, Z. P., Zhong, Y. S., and Pan, J. H. (2022). An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition. *Comput. Biol. Med.* 141:105048. doi: 10.1016/j.compbiomed.2021.105048

Huang, H. P., Hu, Z. C., Wang, W. M., and Wu, M. (2020). Multimodal emotion recognition based on ensemble convolutional neural network. *IEEE Access.* 8, 3265–3271. doi: 10.1109/access.2019.2962085

Huang, Y., Yang, J., Liu, S., and Pan, J. (2019). Combining facial expressions and electroencephalography to enhance emotion recognition. *Fut Internet.* 11:105. doi: 10.3390/fi11050105

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: a database for emotion analysis; using physiological signals. *IEEE Trans Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Kong, W., Zhou, Z., Jiang, B., Babiloni, F., and Borghini, G. J. N. (2017). Assessment of driving fatigue based on intra/inter-region phase synchronization. *Neurocomputing* 219, 474–482. doi: 10.1016/j.neucom.2016.09.057

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis. Comput.* 65, 23–36. doi: 10.1016/j.imavis.2017.02.001

Li, C., Bian, N., Zhao, Z. P., Wang, H. S., and Schuller, W. (2023). Multi-view domain-adaptive representation learning for EEG-based emotion recognition. *Informat Fusion.* 104:102156. doi: 10.1016/j.inffus.2023.102156

Li, Y. J., Huang, J. J., Zhou, H. Y., and Zhong, N. (2017). Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. *Appl. Sci.* 7:20. doi: 10.3390/app7101060

Li, J. P., Zhang, Z. X., and He, H. G. (2018). Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cogn. Comput.* 10, 368–380. doi: 10.1007/s12559-017-9533-x

Liu, Y., Ding, Y. F., Li, C., Cheng, J., Song, R. C., Wan, F., et al. (2020). Multi-channel EEG-based emotion recognition via a multi-level features guided capsule network. *Comput. Biol. Med.* 123:103927. doi: 10.1016/j.compbiomed.2020.103927

Liu, J., Wu, G., Luo, Y., Qiu, S., Yang, S., Li, W., et al. (2020). EEG-based emotion classification using a deep neural network and sparse autoencoder. *Front Syst Neurosci.* 14:14:43. doi: 10.3389/fnsys.2020.00043

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). AffectNet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans Affect. Comput.* 10, 18–31. doi: 10.1109/TAFFC.2017.2740923

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain Comput Interfaces.* 1, 66–84. doi: 10.1080/2326263X.2014.912881

Nguyen, H.-D., Kim, S.-H., Lee, G.-S., Yang, H.-J., Na, I.-S., and Kim, S. H. J. I. T. (2019). Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks.13, 226–237

Peng, G. Q., Zhao, K. Y., Zhang, H., Xu, D., and Kong, X. Z. (2023). Temporal relative transformer encoding cooperating with channel attention for EEG emotion analysis. *Comput. Biol. Med.* 154:106537. doi: 10.1016/j.compbiomed.2023.106537

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2015), FitNets: hints for thin deep nets. arXiv [Preprint].

Sander, K., and Ioannis, P. (2013). Fusion of facial expressions and EEG for implicit affective tagging. *Image Vis. Comput.* 31, 164–174. doi: 10.1016/j.imavis.2012.10.002

Sha, T. H., Zhang, Y. K., Peng, Y., and Kong, W. Z. (2023). Semi-supervised regression with adaptive graph learning for EEG-based emotion recognition. *Math. Biosci. Eng.* 20, 11379–11402. doi: 10.3934/mbe.2023505

Shen, F. Y., Dai, G. J., Lin, G., Zhang, J. H., Kong, W. Z., and Zeng, H. (2020). EEG-based emotion recognition using 3D convolutional recurrent neural network. *Cogn. Neurodyn.* 14, 815–828. doi: 10.1007/s11571-020-09634-1

Siddharth, T. P., and Jung, T. J. S. (2022). Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing, IEEE trans. *Affect. Comput.* 13, 96–107. doi: 10.1109/TAFFC.2019.2916015

Soleymani, M., Asghari-Esfeden, S., Fu, Y., and Pantic, M. (2016). Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Trans Affect. Comput.* 7, 17–28. doi: 10.1109/TAFFC.2015.2436926

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Trans Affect. Comput.* 3, 42–55. doi: 10.1109/T-AFFC.2011.25

Song, T. F., Zheng, W. M., Song, P., and Cui, Z. (2020). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* 11, 532–541. doi: 10.1109/taffc.2018.2817622

Tao, W., Li, C., Song, R. C., Cheng, J., Liu, Y., Wan, F., et al. (2023). EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Trans Affect. Comput.* 14, 382–393. doi: 10.1109/taffc.2020.3025777

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE J Select Topics Signal Process.* 11, 1301–1309. doi: 10.1109/JSTSP.2017.2764438

Wang, Y., Huang, Z., McCane, B., and Neo, P. (2018). EmotioNet: a 3-D convolutional neural network for EEG-based emotion recognition, In: *International Joint Conference on Neural Networks (IJCNN)*. IEEE. pp. 1–7

Wang, D., and Shang, Y. (2013). Modeling physiological data with deep belief networks. *Int J Inf Educ Technol.* 3, 505–511. doi: 10.7763/ijiet.2013.V3.326

Wang, Q., Wang, M., Yang, Y., and Zhang, X. L. (2022). Multi-modal emotion recognition using EEG and speech signals. *Comput. Biol. Med.* 149:105907. doi: 10.1016/j.compbiomed.2022.105907

Wang, Z., Zhou, X., Wang, W., and Liang, C. (2020). Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video. *Int. J. Mach. Learn. Cybern.* 11, 923–934. doi: 10.1007/s13042-019-01056-8

Xue, J., Zhang, T., Chen, P., Philip Chen, C. L., Liu, Z. L., Chen, L., et al. (2020). Multi-Channel EEG based emotion recognition using temporal convolutional network and broad learning system, *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. pp. 2452–2457.

Yang, Y., Wu, Q., Fu, Y., and Chen, X. (2018b). Continuous convolutional neural network with 3D input for EEG-based emotion recognition. In: *Neural Information Processing: 25th International Conference*. ICONIP 2018, Siem Reap, Cambodia, December 13–16, Proceedings, Part VII 25, Springer 2018. pp. 433–443

Yang, Y., Wu, Q., Qiu, M., Wang, Y., and Chen, X. (2018a). Emotion recognition from Multi-Channel EEG through parallel convolutional recurrent neural network. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–7

Yin, Y., Zheng, X., Hu, B., Zhang, Y., and Cui, X. (2021). EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. *Appl. Soft Comput.* 100:106954. doi: 10.1016/j.asoc.2020.106954

Zhang, Y., Cheng, C., Wang, S., and Xia, T. (2022). Emotion recognition using heterogeneous convolutional neural networks combined with multimodal factorized bilinear pooling. *Biomed. Signal Process. Control* 77:103877. doi: 10.1016/j.bspc.2022.103877

Zhang, D., Yao, L., Chen, K., and Monaghan, J. (2019). A convolutional recurrent attention model for subject-independent EEG signal analysis. *IEEE Signal Process Lett.* 26, 715–719. doi: 10.1109/LSP.2019.2906824

Zheng, W. L., and Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zhong, X., Yin, Z., and Zhang, J. (2020), Cross-subject emotion recognition from EEG using convolutional neural networks, In: *2020 39th Chinese Control Conference (CCC)*. pp. 7516–7521