applied sciences

MDPI

Article

Cascaded Convolutional Recurrent Neural Networks for EEG Emotion Recognition Based on Temporal–Frequency–Spatial Features

Yuan Luo [1,2], Changbo Wu [1,2,*] and Caiyun Lv [1]

[1] Key Laboratory of Optoelectronic Information Sensing and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; luoyuan@cqupt.edu.cn (Y.L.); 15155065683@163.com (C.L.)
[2] School of Advanced Manufacturing Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
* Correspondence: 18256220489@163.com

Featured Application: The proposed method can improve emotion recognition accuracy in human–computer interactions.

Abstract: Emotion recognition is a research area that spans multiple disciplines, including computational science, neuroscience, and cognitive psychology. The use of electroencephalogram (EEG) signals in emotion recognition is particularly promising due to their objective and nonartefactual nature. To effectively leverage the spatial information between electrodes, the temporal correlation of EEG sequences, and the various sub-bands of information corresponding to different emotions, we construct a 4D matrix comprising temporal–frequency–spatial features as the input to our proposed hybrid model. This model incorporates a residual network based on depthwise convolution (DC) and pointwise convolution (PC), which not only extracts the spatial–frequency information in the input signal, but also reduces the training parameters. To further improve performance, we apply frequency channel attention networks (FcaNet) to distribute weights to different channel features. Finally, we use a bidirectional long short-term memory network (Bi-LSTM) to learn the temporal information in the sequence in both directions. To highlight the temporal importance of the frame window in the sample, we choose the weighted sum of the hidden layer states at all frame moments as the input to softmax. Our experimental results demonstrate that the proposed method achieves excellent recognition performance. We experimentally validated all proposed methods on the DEAP dataset, which has authoritative status in the EEG emotion recognition domain. The average accuracy achieved was 97.84% for the four binary classifications of valence, arousal, dominance, and liking and 88.46% for the four classifications of high and low valence–arousal recognition.

Keywords: emotion recognition; electroencephalogram; 4D features; convolution; attention

Citation: Luo, Y.; Wu, C.; Lv, C. Cascaded Convolutional Recurrent Neural Networks for EEG Emotion Recognition Based on Temporal–Frequency–Spatial Features. *Appl. Sci.* **2023**, *13*, 6761. https://doi.org/10.3390/app13116761

Academic Editors: Fei Jiang, Aimin Zhou, Ran Wu and Feng Liu

Received: 21 April 2023
Revised: 25 May 2023
Accepted: 26 May 2023
Published: 2 June 2023

1. Introduction

Emotion represents a complex psychological construct that encompasses an individual's affective, cognitive, and behavioural responses to external stimuli, accompanied by corresponding physiological reactions [1]. In the modern era of intelligent technology, people's daily lives are becoming increasingly intertwined with these advanced systems, underscoring the growing importance of accurate emotion recognition in human–computer interaction. Studies in neurological physiology [2] and social psychology [3] have revealed a strong correlation between EEG signals and numerous cognitive processes, including emotional responses, and can objectively reflect the real emotions of subjects, establishing EEG-based emotion recognition as a cutting-edge research area in cognitive science [4].

Owing to the nonlinear, unsmooth, low signal-to-noise ratio (SNR) and multichannel correlation characteristics of EEG data, which are inherently complex chaotic data, there are numerous data processing methods available in the research field [5], but the extraction

of features that are highly relevant to emotions and the selection of a suitable classification model remain two obstacles for researchers. EEG signals have been demonstrated to contain emotion-related feature components in three dimensions: temporal, frequency, and spatial. Unfortunately, most of the literature considers only one or two of these three dimensions and performs pattern recognition by simple combinations, completely ignoring spatial information interactions across channels, prior knowledge across frequency bands, and information complementarity between different features. This not only fails to improve model accuracy, but also leads to feature redundancy and increased complexity. Therefore, it is important to improve emotion recognition accuracy by integrating information from different domains and adaptively capturing important temporal, frequency, and spatial features in subsequent classification models.

Recently, deep learning, the successor to machine learning, has made significant contributions to various fields, including computer vision (CV) and natural language processing (NLP). A wide range of deep models have been developed and have shown impressive results. The application of these models has also had a significant impact on EEG emotion recognition, including the use of autoencoders (AEs), graph convolutional neural networks (GCNs), transformers, and other related techniques. Although these approaches have yielded promising outcomes, challenges remain that need to be addressed. Specifically, how to construct a network that is suitable for EEG signal feature extraction represents a crucial issue that must be addressed.

Regarding the aforementioned challenges, a cascade network based on multidimensional features is proposed, which presents three main contributions:

- EEG data are converted into a 4D matrix structure consisting of multiple frames, which contain information in three dimensions: temporal, frequency, and spatial, and can effectively represent the neural features of different emotions.
- In this paper, a novel attention module FcaNet is introduced. FcaNet redistributes the weights of different channels to obtain high-quality discrimination. FcaNet is found to be superior to traditional channel attention squeeze-and-excitation networks (SENet) while incurring no significant computational cost.
- To satisfy the real-time demands of the emotion recognition system, a residual network is devised, which comprises DC and PC to decrease the computational burden while utilizing the attributes of depth-separable convolution to segregate the spatial and channel mixing dimensions. Furthermore, the existence of a residual structure prevents overfitting. Ultimately, Bi-LSTM is employed to understand the temporal interdependence among different frames in the sample. The hidden layer states at each frame moment are allocated weights and then summed to serve as the input to softmax.

The experimental results show that the designed model achieves advanced performance on the DEAP dataset. The rest of the article is arranged as follows: Section 2: Related Work, Section 3: Methods, Section 4: Materials and Experimental Results, and Section 5: Conclusions.

2. Related Work

In the realm of emotion recognition, machine learning has traditionally been a popular approach for simple classification tasks [6]. Some prominent algorithms, including support vector machine (SVM), decision trees (DT), and k-nearest neighbour (KNN), have been successfully utilized in this field. Zubair [7] used the discrete wavelet transform (DWT) to extract temporal–frequency information and applied the maximum relevancy and minimum redundancy algorithm (mRMR) to select the most relevant features. In the literature [8,9], wavelet transform (WT) was applied for sub-band EEG signal decomposition, and the processed smooth feature information was then fed into SVM for classification. This method demonstrated promising improvement in accuracy for EEG emotion-state recognition in machine learning. However, machine learning techniques still face signifi-

cant limitations when processing nonlinear and indistinguishable data, which restrict their capability for more complex classification tasks.

With the advent of deep learning, the machine learning limitations are gradually being overcome, and deep learning techniques are successfully moving in the direction of EEG emotions. Typically, deep-learning-based approaches focus on feature extraction from three dimensions: temporal, frequency, and spatial. For instance, in terms of the temporal information of EEG signals, Xing [10] proposed a framework combining a stacked autoencoder (SAE) and long short-term memory neural networks (LSTM). SAE is employed to simulate the mixing process in EEG and to separate the source signals. Then, the source signals are framed, and frequency features are extracted and combined into chained data, followed by discriminative classification using LSTM. Ma [11] designed the multimodal residual LSTM (MMResLSTM) network, using different LSTM layers to learn the temporal characteristics of different physiological signals and share parameters to achieve the information interaction of different modal data. References in the literature [12–14] introduced a temporal learning architecture that employs a 1D convolutional neural network (CNN) to extract temporal information from multichannel chained data.

Studies in neuroscience and psychology have proven that the δ-band (1–4 Hz), θ-band (4–8 Hz), α-band (8–13 Hz), β-band (13–30 Hz), and γ-band (30–50 Hz) of EEG [15] are associated with human emotions. Therefore, during the frequency domain feature extraction process, EEG signals are typically mapped onto these five frequency bands, and sub-band features are extracted. Zheng [16] extracted EEG signal differential entropy (DE) features on a sub-band and performed emotion recognition on a deep belief network (DBN). Zhang [17] extracted DE, power spectral density (PSD), differential asymmetry (DASM), and rational asymmetry (RASM) in four representative EEG datasets and trained them for classification by the proposed dynamic graph convolutional neural network (DGCNN). The results showed that the DE and PSD features could convey the most discriminative information.

Researchers have been interested in exploring physical models between electrode positions to characterize spatial features in EEG signals. Hwang [18] proposed a method to generate an image by performing a polar coordinate projection of the channel DE features and using different interpolation methods to fill the blank space after the projection, thus proving that the spatial topology based on electrode arrangement is effective. Song [19] utilized graph theoretic ideas to model multichannel EEG signals and used DGCNN to explore the depth spatial information of neighbouring channels. Other studies [20–23] constructed a connectivity matrix containing structural information of the brain to express features in different ways and then input the rearranged EEG signals into an end-to-end CNN model.

The researchers have also considered various feature information combinations. References from the literature [23,24] extracted the DE features of the four EEG signal sub-bands, which were mapped into a 3D matrix based on the electrode distribution to retain its channel information. Finally, the spatial–frequency information was extracted by different 2D convolutions. Researchers in [25–28] introduced a combined CNN and LSTM model that learns spatial–frequency and temporal features, respectively, from the input signal. Experimental findings reveal that the accuracy of combined multidimensional feature information surpasses that of a single dimension.

There exists a broad range of feature extraction methods; however, fully exploiting key features remains a significant challenge. Introducing an attention mechanism has greatly enhanced the capabilities of various classification models. Researchers in EEG emotion recognition have noted that attentional mechanisms can selectively focus on brain regions associated with emotional stimulation and have begun to explore their application to EEG emotion recognition to improve performance. Zhang [29] introduced band attention and temporal attention in a hybrid deep learning model to adaptively assign weights for different frequency bands and times, respectively. In [22,30], researchers constructed 3D

data containing temporal–frequency information, introduced the channel attention module SENet to assign weights to frequency bands and channels, and obtained advanced results.

3. Methods

In this section, data preprocessing and the three components of the proposed model are explained in detail. A complete overview of the model framework is shown in Figure 1.

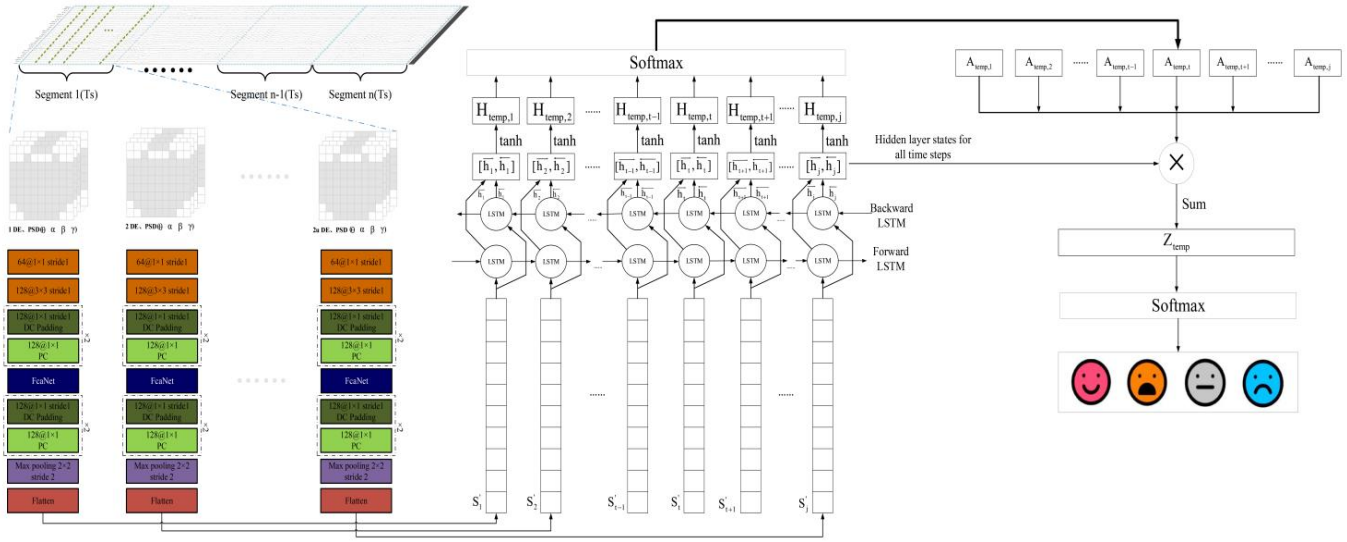


Figure 1. Overview of the identification framework.

3.1. Data Preprocessing

In the original data acquisition, a few seconds of the steady state of the subject before the emotional stimulus is recorded. Most studies choose to remove this part of the data directly when recognizing the task and consider only the EEG signal in the stimulated state. Studies [25,30,31] demonstrated that preprocessing through the baseline signal is effective for improving experimental robustness. The difference between the signal of the subject under the emotional stimulus and the baseline signal was calculated as an indication that the segment's emotional state data could yield the expected results. Assume that the dataset is sampled at a frequency of S. Consider the entire baseline clip $X_{base} \in R^{MxN_1}$, where M is the number of electrodes: 32; and N_1 is the sum of the sampling points for the entire baseline clip. First, X_{base} is uniformly divided into 1 s periods to obtain $\left\{X^1_{base}, X^2_{base}, \ldots, X^O_{base}\right\}$, $X^i_{base} \in R^{MxS} (i = 1, 2, \ldots, O)$ denoting the i-th baseline segment, $O = \frac{N_1}{S}$. Then, the average value $\overline{X}_{base} \in R^{MxS}$ of X_{base} is calculated as follows:

$$\overline{X}_{base} = \frac{\sum_{i=1}^{O} X^i_{base}}{O} \tag{1}$$

$X_{trial} \in R^{MxN_2}$ denotes the experimental EEG signal in the emotional stimulation condition, and N_2 denotes the total number of sampling points of the experimental EEG signal. As in the above method, the experimental data were split into Q segments according to the length criterion of $\overline{X}_{base}$ to obtain $\left\{X^1_{trial}, X^2_{trial}, \ldots, X^Q_{trial}\right\}$, where $X^i_{trial} \in R^{MxS}$ $(i = 1, 2, \ldots, Q)$ denotes the i-th experimental segment, and $Q = \frac{N_2}{S}$. Finally, the baseline removal data were obtained by subtracting the baseline mean $\overline{X}_{base}$ from each segment of the experimental data using the following equation:

$$X^i_{trial.rmov} = X^i_{trial} - \overline{X}_{base}, X^i_{trial.rmov} \in R^{M \times S} \tag{2}$$

where $X^{i}_{trial.rmov}(i = 1, 2, \ldots, Q)$ represents the contrast data from X^{i}_{trial} and $\overline{X}_{base}$. Finally, the Q segment removed the baseline data to patch the complete data $X_{trial.rmov} = \left\{X^{1}_{trial.rmov}, X^{2}_{trial.rmov}, \ldots, X^{Q}_{trial.rmov}\right\}$, $X_{trial.rmov} \in R^{MxN_2}$, which was performed on $X_{trial.rmov}$ in the subsequent feature extraction.

3.2. Multiband Four-Dimensional Feature Construction

In this section, the process of forming a 4D matrix based on the temporal–frequency–spatial is explained in detail, as shown in Figure 2.

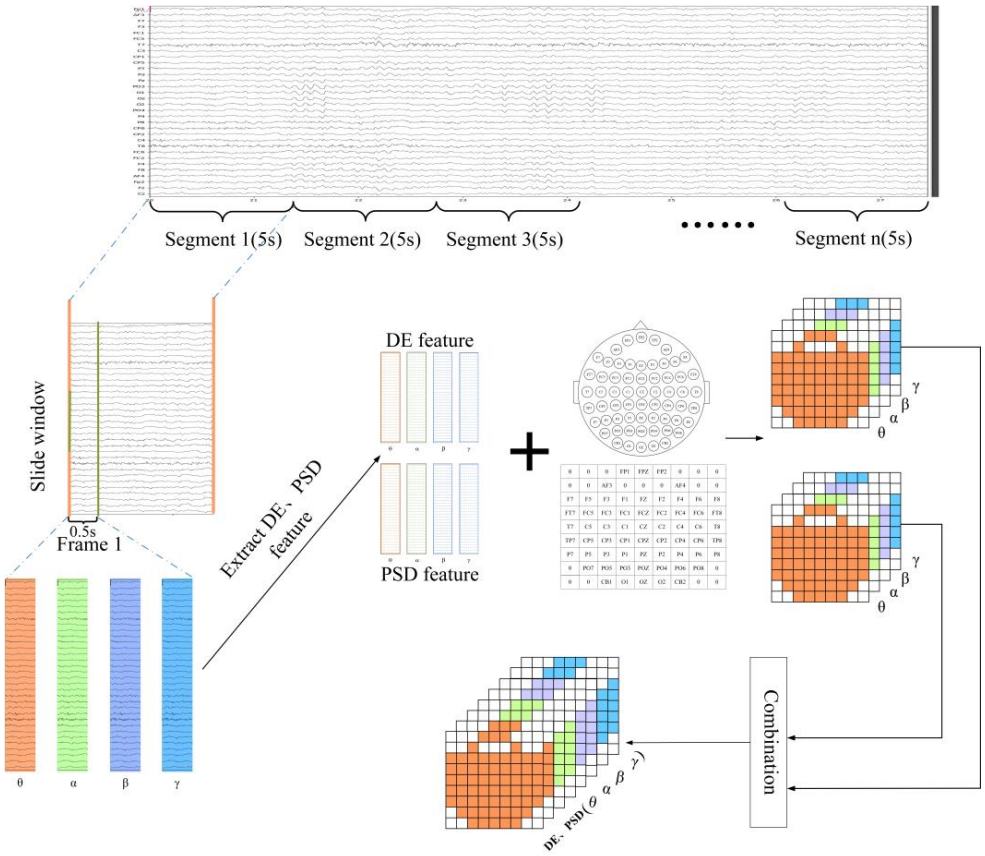


Figure 2. 4D feature extraction process.

Since the integration of temporal and frequency band information of all available channels requires a three-dimensional spatial representation, this requires the consideration of a matrix of filtered signals obtained from all channels. Take the DEAP dataset as an example, which includes a total of 1280 sample data points (32 subjects × 40 experiments). For the problem that a small quantity of data can cause overfitting when the network layers are deep, we reduce this effect by increasing the number of samples. Specifically, for the signal $X_{trial.rmov} \in R^{M \times N_2}$ after baseline removal, a nonoverlapping windowing process with a time duration of u seconds (design in this paper with u = 5 s) is performed to obtain the divided data $\left\{X^{1}_{S}, X^{2}_{S}, \ldots, X^{n}_{S}\right\}$, where the ith window segment $X^{i}_{S} = \{x_1, x_2, \ldots, x_M\} \in R^{M \times R}(i = 1, 2, \ldots, n)$, $x_c \in R^{R}(c = 1, 2, \ldots, M)$ denotes data from the c-th electrode channel in X^{i}_{S}, and R is the sampling frequency of the whole window segment.

Different frequency band data in EEG reflect different emotional information, as detailed in Table 1 [32]. While conducting directional research, the linear data phase is commonly required to be high. A finite impulse response filter (FIR) can obtain a rigorous linear phase with high stationarity and low interference power caused by operational errors. Therefore, we use FIR to have a sub-band division for each window segment X^{i}_{S}.

Qu [25] experimentally compared the EEG emotion recognition results under different band combinations and found that band combinations of α, β, and γ had the highest accuracy in task recognition. In addition, Frantzidis [33] remarked that the θ-band features are closely correlated with arousal. Therefore, in this paper, the four bands $\theta, \alpha, \beta, \gamma$ were chosen to study the emotional state features of EEG signals. Through the FIR filter, each window segment was decomposed into $X_S^i = \left\{\left\{f_1^\theta, f_1^\alpha, f_1^\beta, f_1^\gamma\right\}, \ldots, \left\{f_M^\theta, f_M^\alpha, f_M^\beta, f_M^\gamma\right\}\right\}$ $(i = 1, 2, \ldots, n)$. The specific realization formula is as follows:

$$h(n) = h_d(n) \cdot w(n) = \frac{\sin[(M-\tau)\cdot W_h] - \sin[(M-\tau)\cdot W_l]}{\pi \cdot (m-\tau)} \cdot w(n) \tag{3}$$

$$H_d(e^{jw}) = \sum_{n=-\infty}^{n=+\infty} h_d(n) e^{-jwn} \tag{4}$$

$$H(w) = |H(e^{jw})| = |\frac{1}{2\pi} \int_{-\pi}^{\pi} (H_d(e^{j\theta}) e^{-j\theta M/2}) W(e^{jw}) d\theta| \tag{5}$$

where $h(n)$ is the filter coefficient, $H_d(e^{jw})$ is the corresponding frequency response, $H(w)$ is the amplitude–frequency response function, $h_d(n)$ is the unit impulse response, $w(n)$ is the window function, W_h and W_l are the cut-off frequencies of the bandpass filter, $\tau = \frac{M-1}{2}$, and M is the number of filter steps.

Table 1. Emotional states correspond to different frequency band information.

Frequency Band	Frequency Range	Brain States	Awareness
δ	1–4	Extreme fatigue and deep sleep states	Sleep mode
θ	4–8	Light sleep, frustrated state	Low
α	8–13	Awake, quiet, and eyes closed state	Medium
β	13–30	Active thinking, mental tension, anxiety, concentration state	High
γ	30–50	Multimodal sensory stimulation, mentally active state	High

Treating all channel data of a single frequency band as a whole, such that X_S^i is transformed into $\left\{p_i^\theta, p_i^\alpha, p_i^\beta, p_i^\gamma\right\}$, where $p = \{p_1, p_2, \ldots, p_M\}$ denotes all channel data of a single frequency band, the arrangement order is shown in Figure 3.

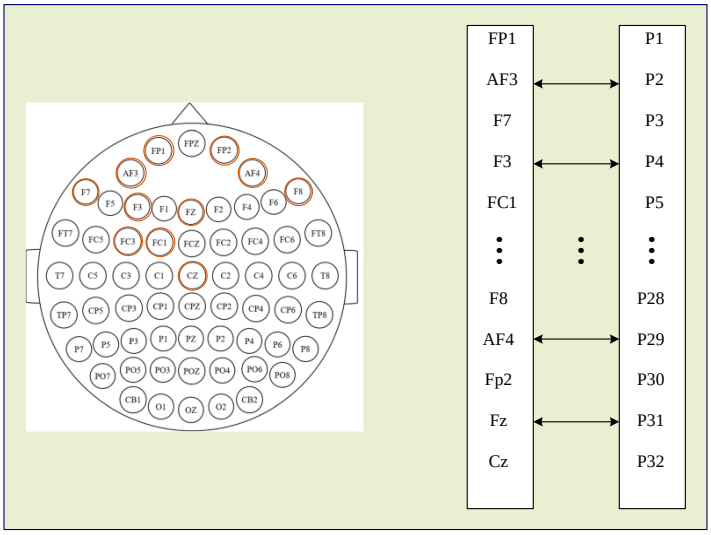


Figure 3. Channel alignment order.

After data enhancement, our focus shifts to emotion recognition at the segmentation level, and considering that human emotion changes are temporally dynamic, the window segmented signal X_S^i is segmented into equal-length frames of 0.5 s. Using 0.5 s data per channel as the vector components that constitute a single frame window, X_S^i will be converted into a sequence containing 2u frame vectors, $X_S^i = \left\{\left\{p_i^{1\theta}, p_i^{1\alpha}, p_i^{1\beta}, p_i^{1\gamma}\right\}, \ldots, \left\{p_i^{j\theta}, p_i^{j\alpha}, p_i^{j\beta}, p_i^{j\gamma}\right\}\right\}$ $(i = 1, 2, \ldots, n.j = 1, 2, \ldots, 2u)$. Considering the information complementarity between different features, the DE and PSD features of all channels within each frame window are extracted in this paper [16].

DE is a derivative of Shannon's concept of information entropy over a continuous probability distribution, which is a good method to describe internal EEG information. A specific length of EEG that approximately obeys the Gauss distribution $N(\mu, \sigma_i^2)$ is calculated as:

$$DE = -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}}\right) dx = \frac{1}{2}\log(2\pi e\sigma_i^2) \tag{6}$$

where σ is the variance in the sequence signal.

PSD is a physical quantity that characterizes the relationship between the power energy of a signal and its frequency and is often used to study random vibration signals, which can describe the activation level and emotional complexity of EEG signals. It is calculated as:

$$P(w) = \frac{1}{MU}\left|\sum_{n=0}^{M-1} x_N^i(n) d(n) e^{-jwn}\right|^2 \tag{7}$$

where $x_N^i(n)$ is the sampled data in segment i, $d(n)$ is the selected window function, M is the length of each segment, and U is the normalization factor.

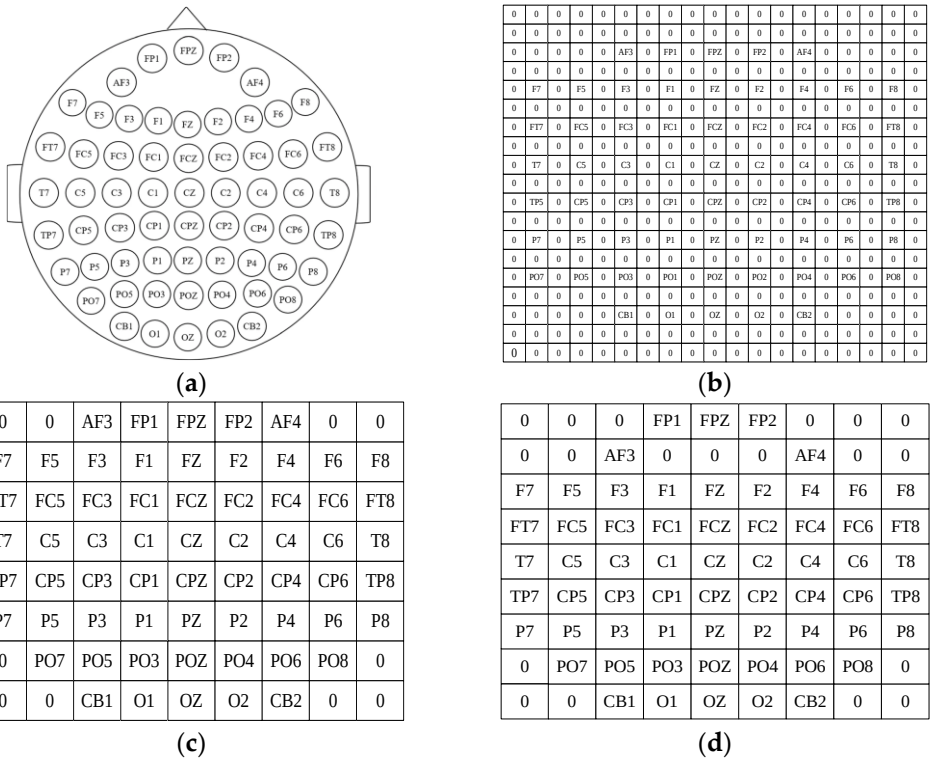
The final two sets of features obtained both include a two-dimensional vector sequence of four frequency bands, where the frame vectors $f_i^{j\gamma}$ and $g_i^{j\gamma}$ retain the same arrangement as $p_i^{j\gamma}$. It can be described as:

$$F_{PSD} = PSD(X_S^i) = \left\{\left\{f_i^{1\theta}, f_i^{1\alpha}, f_i^{1\beta}, f_i^{1\gamma}\right\}, \ldots, \left\{f_i^{j\theta}, f_i^{j\alpha}, f_i^{j\beta}, f_i^{j\gamma}\right\}\right\} \tag{8}$$

$$F_{DE} = DE(X_S^i) = \left\{\left\{g_i^{1\theta}, g_i^{1\alpha}, g_i^{1\beta}, g_i^{1\gamma}\right\}, \ldots, \left\{g_i^{j\theta}, g_i^{j\alpha}, g_i^{j\beta}, g_i^{j\gamma}\right\}\right\} \tag{9}$$

Nevertheless, frequency features can never fully characterize all the feature components of the entire signal. The EEG is obtained from electrodes placed in different regions for acquisition, and the positional relationships between these electrodes contain information about the spatial structure associated with emotions. Therefore, in this paper, a single frequency band frame vector is mapped into a two-dimensional matrix using spatial mapping. References [34–36] proposed the sparse transform, compact transform, and sensitive transform methods, respectively. The sparse transformation matrix is 19 × 19, which undoubtedly requires many computations [34]. The compact matrix of Shen [35] reduces the size to 8 × 9, which drastically reduces the computational effort, and the adjacent electrodes in the matrix are more strongly connected, but it is not sensitive to spatial information and does not work well experimentally. Therefore, here, we use the sensitive transformation method of Xu [36] to map it to 9 × 9 data. Compared with the first two methods, the connection relationship between electrode points is more in line with the "10/20" system while considering the computational effort. These three specific mapping methods are shown in Figure 4. Each frame vector $f_i^{j\gamma}$ of a single frequency band is further converted into a two-dimensional matrix. SEED and DEAP each contain 62 and 32 electrode channels, so here, 62 channels are used to map the electrode positions, the corresponding DE and PSD values are filled for the elements with position mapping, and the remaining positions

are replaced by 0. In this way, the window segments transform into a two-dimensional matrix sequence from the perspective of a single frequency band. Taking the θ-band as an example, $S^{\theta} = \left\{S_1^{\theta}, S_2^{\theta}, \ldots, S_j^{\theta}\right\} \in R^{9\times9\times2u}(j = 1, 2, \ldots, 2u)$, it is extended into 3 dimensions in dimensionality. The four-dimensional data of a single feature in the whole window fragment are obtained by fusing the four frequency bands. Taking the DE feature of the window fragment as an example, $S = \left\{S_{(DE)1}, S_{(DE)2}, \ldots, S_{(DE)j}\right\} \in R^{9\times9\times8\times2u}$, considering the problem of information complementarity between different features, we place the same frequency band data of DE and PSD features together to construct the final four-dimensional data $S = \{S_1, S_2, \ldots, S_j\} \in R^{9\times9\times8\times2u}$.



0	0	AF3	FP1	FPZ	FP2	AF4	0	0
F7	F5	F3	F1	FZ	F2	F4	F6	F8
FT7	FC5	FC3	FC1	FCZ	FC2	FC4	FC6	FT8
T7	C5	C3	C1	CZ	C2	C4	C6	T8
TP7	CP5	CP3	CP1	CPZ	CP2	CP4	CP6	TP8
P7	P5	P3	P1	PZ	P2	P4	P6	P8
0	PO7	PO5	PO3	POZ	PO4	PO6	PO8	0
0	0	CB1	O1	OZ	O2	CB2	0	0

0	0	0	FP1	FPZ	FP2	0	0	0
0	0	AF3	0	0	0	AF4	0	0
F7	F5	F3	F1	FZ	F2	F4	F6	F8
FT7	FC5	FC3	FC1	FCZ	FC2	FC4	FC6	FT8
T7	C5	C3	C1	CZ	C2	C4	C6	T8
TP7	CP5	CP3	CP1	CPZ	CP2	CP4	CP6	TP8
P7	P5	P3	P1	PZ	P2	P4	P6	P8
0	PO7	PO5	PO3	POZ	PO4	PO6	PO8	0
0	0	CB1	O1	OZ	O2	CB2	0	0

Figure 4. (**a**) EEG; (**b**) sparse transform; (**c**) compact transform; and (**d**) sensitive transform.

After obtaining the mapping matrix of all frame window feature data, Equation (10) is used to normalize the data to a distribution with a mean of 0 and a variance of 1, attempting to avoid poles that may negatively affect the recognition results during gradient descent.

$$x_{scale} = \frac{(x - x_{mean})}{\sigma} \tag{10}$$

where x_{mean} is the mean of the eigenvalue data, σ is the standard deviation of each set of features, x is the actual eigenvalue, and x_{scale} is the final normalized data.

3.3. Network Architecture

3.3.1. FcaNet

In this study, the aim is to obtain high-quality EEG feature information. To achieve this goal, we incorporated an attention module called FcaNet [37] into the backbone network, which is used to lower the weight of low-quality EEG information. FcaNet is a novel attention module based on SENet [38], which was initially used in target detection. Unlike the global average pooling (GAP) used in SENet to squeeze the feature map, FcaNet uses two-dimensional discrete cosine transformation (2D-DCT) for the same purpose. GAP corresponds only to the lowest frequency portion of the 2D-DCT, resulting in the loss of the

remaining frequency portions in the channel. A comparison of SENet and FcaNet is shown in Figure 5.

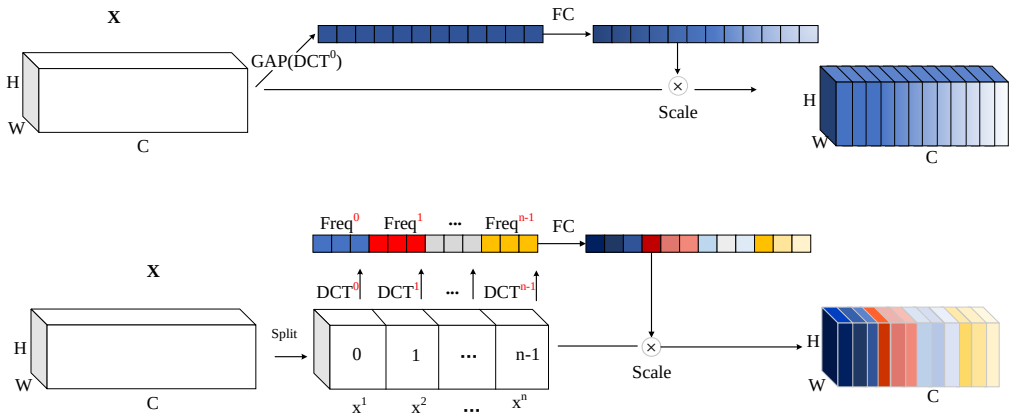


Figure 5. Comparison of SENet and FcaNet.

Given an input X, it is divided into n parts along the channel dimension: $\left[X^0, X^1, X^2, \ldots, X^{n-1}\right]$, where $i \in \{0, 1, 2, \ldots, n-1\}$, $X^i \in R^{H \times W \times C'}$, $C' = \frac{C}{n}$. For each part, the corresponding 2D-DCT frequency portion is distributed, and the result of 2D-DCT is used as the compression result noted by the channel. The following equation shows 2D-DCT:

$$Freq^i = 2DDCT^{u_i, v_i}(X^i) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X^i_{:,h,w} B^{u_i, v_i}_{h,w} \tag{11}$$

$$B^{u_i, v_i}_{h,w} = \cos\left(\frac{\pi h}{H}\left(u_i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(v_i + \frac{1}{2}\right)\right) \tag{12}$$

$$s, t.i \in \{0, 1, \ldots, n-1\}$$

where H, W, and (u_i, v_i) are the height, width, and 2D index of X^i, respectively. The whole feature information compression vector can be represented by cascading as:

$$Freq = compress(X) = cat([Freq^0, Freq^1, \ldots, Freq^{n-1}]) \tag{13}$$

The complete FcaNet framework can be described as follows:

$$ms\_att = sigmoid(fc(Freq)) \tag{14}$$

3.3.2. Spatial–Frequency Feature Learning

Extracting spatial–frequency features is primarily accomplished through the use of the convolutional encoder and the attention module FcaNet discussed in Section 3.3.1, with the module structure illustrated in Figure 6.

Specifically, 2u frames S_j in each segment are fed into the CNN module sequentially in time order, and the three-dimensional data structure of each frame is $9 \times 9 \times 8$. To better retain information in the relatively small three-dimensional data structure, two convolutional layers are first implemented. The first layer employs a 1×1 convolution kernel with 64 kernels, and the second layer utilizes a 3×3 convolutional kernel with 128 kernels. These different convolutional kernels are utilized to extract deeper information from the three-dimensional data. Subsequently, the residual network is constructed using DC and PC. This combination reduces the number of training parameters while extracting internal features from expanded individual feature maps using DC and expressing cross-feature map relationships using PC. Moreover, the residual structure effectively avoids network degradation. ReLU is utilized as the activation function for each convolutional layer, and BatchNorm processing is performed. Padding operations are carried out for DC

to maintain output size consistency for each layer of convolution. FcaNet assigns weights to different channel features to enhance model performance. The data are subjected to dimensionality reduction through the utilization of a 2 × 2 maximum pooling layer at the end of the cycle, followed by one-dimensional data transformation via flattened layer processing. Finally, each data frame is convolutionally encoded to obtain vector $S'_j \in R^{1152}$.

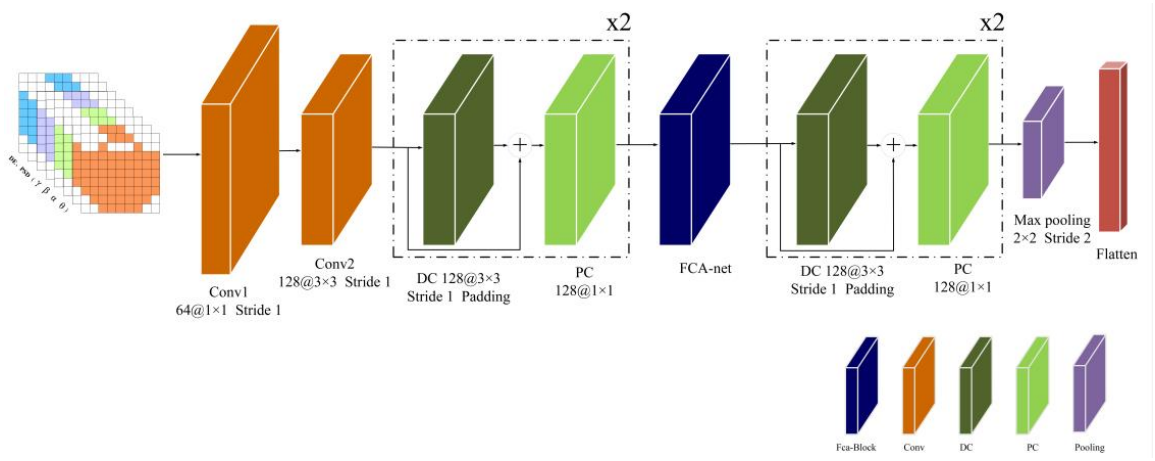


Figure 6. Spatial–frequency feature extraction module.

3.3.3. Temporal Feature Learning

Since emotion changes are temporally dynamic, changes between frames in four-dimensional data may hide emotion-related information. To explore the temporal correlation features in the whole window segment, we input the spatial–frequency features $S' = \left\{S'_1, S'_2, \ldots, S'_j\right\}$ obtained after convolutional coding into the Bi-LSTM network in time order for coding.

LSTM solves the problem that its distant text information cannot be exploited and its close distance but not much semantic association is based on the traditional recurrent neural networks (RNN). It is a model for processing sequential signals that can mitigate the gradient disappearance that occurs with long sequence inputs in RNN. The computational equation of an LSTM cell is shown as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t]) + b_f \tag{15}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t]) + b_i \tag{16}$$

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t]) + b_c \tag{17}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \tag{18}$$

$$O_t = \sigma(W_O \cdot [h_{t-1}, x_t]) + b_O \tag{19}$$

$$h_t = \tanh(C_t) \times O_t \tag{20}$$

where x_t is the current input feature, W and b are the matrix and bias vector to be trained, respectively, i_t, f_t, and O_t are the three gates introduced by LSTM, which are the input gate, forget gate, and output gate, respectively, and the three gates via the sigmoid function, so that the threshold range is controlled between 0 and 1. Cell state C_t characterizes long-term memory. Candidate state $\widetilde{C}_t$ represents the new information to be deposited into C_t by

induction, and h_t is the hidden state. In comparison, the output equation of the Bi-LSTM is shown below:

$$y_t = \sigma(W_h \cdot [h_t, h'_t]) + b_h \tag{21}$$

Bi-LSTM networks combine an LSTM network that moves from the beginning of the sequence and an LSTM network that moves from the end of the sequence, and the backward layer is an extension of the information of past emotions. It is worth mentioning that the parameters of the two LSTM neural networks in Bi-LSTM are mutually independent, and they share only the input vector sequence. This structure merges the gate control architecture and the bidirectional feature perfectly and experimentally proves to be more efficient than a single LSTM for feature extraction of sequences. The network architecture of Bi-LSTM is shown in Figure 7. The time sequences are input to the model, and the forward layer has the information at time t and the previous time, while the backward layer has the information at time t and the subsequent time. The hidden layer data of the two LSTM layers can be processed using summation, average, or connection. Equation (21) is the output value in the form of a connection.

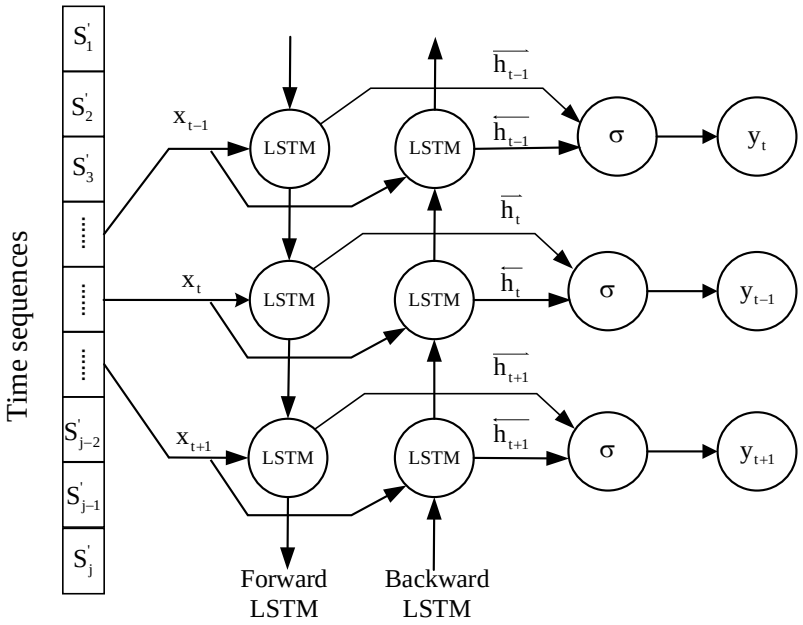


Figure 7. Bi-LSTM structure information.

To highlight the data at key frame moments, the output of Bi-LSTM is not used as the result of the whole temporal feature learning module in this paper. Instead, a nonlinear transformation is first performed for each frame moment hidden layer state with the following equation:

$$H_{temp,j} = \tanh(W_{temp,j} h_j + b_{temp,j}) \tag{22}$$

The amount of memory in each LSTM layer is $\frac{q}{2}$. The hidden layer is processed in a connected form to obtain $h_j \in R^q$. $H_{temp,j}$ is the nonlinear expression of the hidden layer, while $W_{temp,j} \in R^{d \times q}$ and $b_{temp,j} \in R^d$ are the weights and bias vectors of the tanh function, respectively, and d is set to 512. After obtaining $H_{temp,j} \in R^d$, the softmax function is used to calculate the weights for each frame moment to obtain $A_{temp,j}$. The specific equation is as follows:

$$A_{temp,j} = \frac{\exp(H_{temp,j} u_{temp,j})}{\sum_j \exp(H_{temp,j} u_{temp,j})} \tag{23}$$

where $u_{temp,j} \in R^d$ is a trainable parameter, and the greater the value of $A_{temp,j}$, the more important the corresponding frame is in the timing sequence. Multiplying the hidden layer states of all frame data with the weights and summing up, the equation is as follows:

$$Z_{temp} = \sum_{j=1}^{2u} A_{temp,j} h_j \quad (24)$$

where Z_{temp} is used as the output of the whole temporal feature learning model, which not only contains the temporal correlation of the whole window segment, but also enhances the important frame data and suppresses the irrelevant information. Finally, Z_{temp} is used to obtain the prediction result by the softmax classifier. The entire temporal information extraction structure is shown in Figure 8:

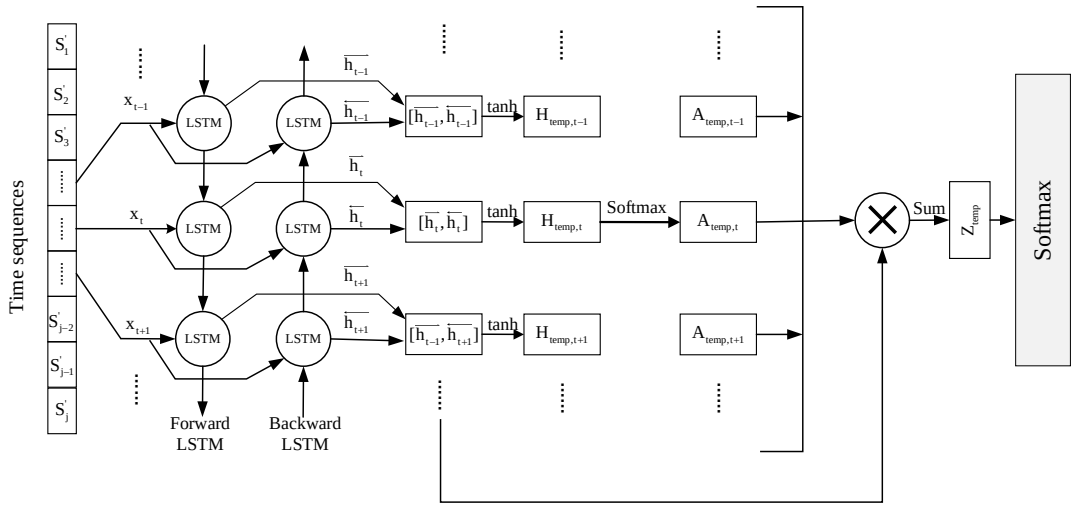


Figure 8. Time feature extraction module.

In the proposed method, we conducted ablation experiments on how many memory cells are set in the unidirectional LSTM layer, and the results are shown in Table 2. Therefore, the unidirectional LSTM layer is set up with 256 memory cells (512 in total) in the proposed method.

Table 2. Experimental comparison of different numbers of memories.

Number	Test Accuracy (%) (2-Class)	F1-Score	Test Accuracy (%) (4-Class)
64	95.09	93.47	86.33
128	95.42	94.35	87.27
256	97.84	96.61	88.46
512	96.57	94.66	87.65

4. Materials and Experimental Results

4.1. Dataset

The DEAP dataset comprises EEG signals from 32 participants, and the acquisition process is illustrated in Figure 9. During data collection, the "10–20" international standard 32-lead electrode cap is used to record signals, and each participant watches 40 one-minute videos while EEG signals are recorded for 63 s (3 s baseline + 60 s of video stimulation) per sample. Thus, the entire dataset consists of 1280 (32 × 40) samples, with each sample containing 63 s of data from 32 channels. Following video viewing, participants subjectively evaluated the videos based on arousal, valence, dominance, and liking using a 1–9 scale. Two versions of the DEAP dataset are officially available: one is the raw signal containing noise such as electromyography (EMG) and electrooculogram (EOG); the other is the

preprocessed data, which consists of downsampling the data to 128 Hz from 512 Hz, filtering and denoising using a 4–45 Hz bandpass filter. This study utilizes the preprocessed version of the DEAP dataset in Python to conduct the experiment. The downsampling operation considerably reduces the computational effort, and the resulting accuracy impact is minor.

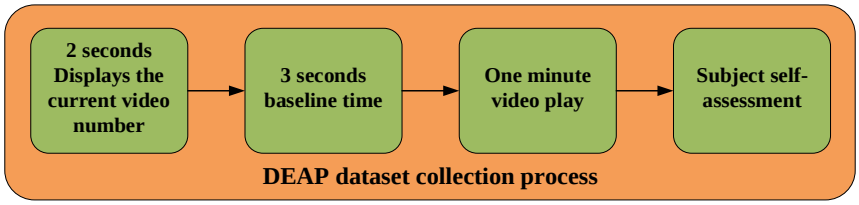


Figure 9. DEAP dataset acquisition process.

Upon removing the baseline from the dataset's signal, a single channel of 60 s will result in 60 × 128 samples, and the resulting $X_{trial.rmov}$ obtained from the 32 channels for a subject will be of the form 32 × 60 × 128. To create more samples, the data were segmented in this study into nonoverlapping windows with a duration of u = 5 s. Specifically, the data were partitioned into 12 windows of 5 s duration, and the data form of X_S^i was 32 × 5 × 128. The follow-ups were all performed with each window fragment X_S^i as a single experimental sample for emotion recognition. The data form is expanded to 32 × 5 × 128 × 4 by dividing X_S^i into frequency bands. To capture the hidden temporal messages of the sequence signal, the single-band data $\left\{p_i^\theta, p_i^\alpha, p_i^\beta, p_i^\gamma\right\}$ in X_S^i are separated into frames of 0.5 s, and each frame has the data form 32 × 0.5 × 128 × 4. The vector data are converted into a two-dimensional matrix with a sensitive transformation on the basis of a frame, and the DE and PSD features of every 0.5 s frame data are used as matrix element data. Each frame of a single frequency band becomes a 9 × 9 matrix, combining two groups of features as well as four frequency bands with different features to obtain the final frame data in the form of 9 × 9 × 8. The whole sample X_S^i is transformed into a four-dimensional feature sequence of 9 × 9 × 8 × 10 when viewed from the perspective of the whole sample X_S^i. Each subject has 40 (video) × 60 (seconds) of EEG data, and using 5 s as a sample, 480 copies (40 × 12) are generated. For 32 subjects, a total of 15,360 samples are generated.

4.2. Experimental Parameter Settings and Evaluation Indices

The code implementation in this article was partially performed on a CUDA 11.2, PyTorch version 1.11 framework, and the hardware module was a server with four Nvidia RTX2080Ti processors, manufactured by Lenovo, Beijing, China. The loss function utilized was cross-entropy, with L2 regularization applied to enhance generalization. The Adam optimizer was used, with a learning rate of 0.001 and a dropout rate of 0.5. A 10-fold cross-validation method was used. To determine the optimal number of iterations for subsequent experiments, a range of values was assessed, and epoch = 50 was chosen based on the achieved accuracy of the model. The outcomes of the comparative analysis are presented in Table 3.

Table 3. Experimental comparison for different epoch cases.

Epoch	Test Accuracy (%) (2-Class)	F1-Score	Test Accuracy (%) (4-Class)
50	97.84	96.61	88.46
100	97.53	96.49	86.93
150	96.81	95.12	86.10
200	97.04	96.83	85.79

Various performance metrics can be employed to assess the performance of a system. Although accuracy is widely used as the evaluation criterion to indicate the percentage of

correctly predicted samples, it is not always the most appropriate metric, particularly in the case of imbalanced data. In this regard, the F1-Score is used as an additional performance metric in this study. The F1-Score, which is the harmonic mean of precision and recall, is a statistical measure that assesses the accuracy of a binary classifier. Specifically, the equation for F1-Score is expressed as follows:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{25}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{26}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{27}$$

$$\mathrm{F-score} = \frac{2 \times \mathrm{Pre} \times \mathrm{Rec}}{\mathrm{Rec} + \mathrm{Pre}} \tag{28}$$

where TP and TN indicate that the data labels are positive and negative classes, respectively, and are consistent with the recognition results, and FP and FN indicate that the data labels are positive and negative classes, respectively, but are inconsistent with the recognition results.

4.3. Emotion Recognition Binary Classification Experiment

To verify the efficacy of the feature extraction method proposed in this article, we selected EEG features that lacked frequency feature extraction and spatial transformation and compared them to the method outlined in this paper. For each set of features, we performed two- and four-classification tasks on the dataset. The set of features lacking frequency feature extraction and spatial transformation is referred to as the "baseline features" throughout this study, the detailed extraction process is shown in Figure 10.

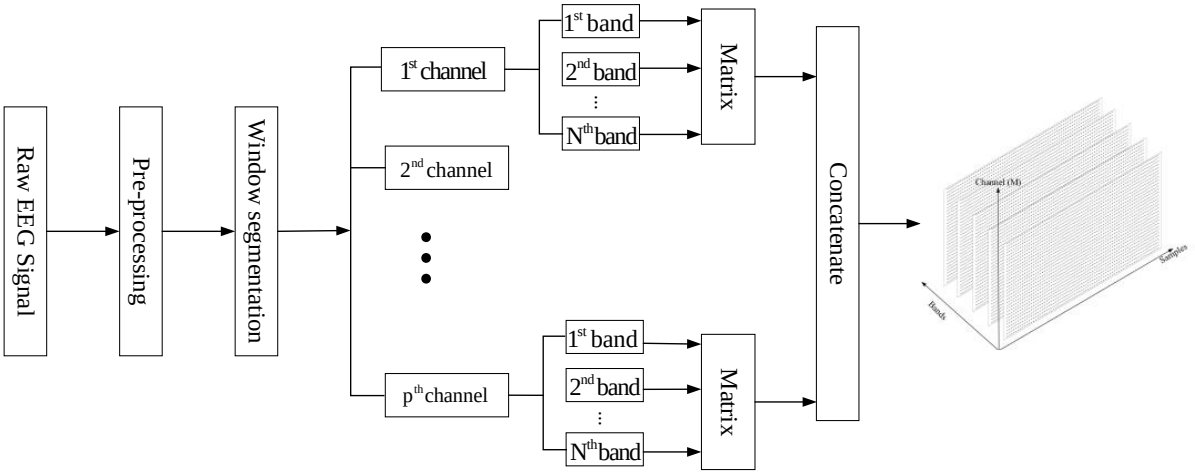


Figure 10. Baseline feature extraction process.

To ensure the maximum effectiveness of the proposed method, the same baseline removal, window segmentation, and frame-segmentation processes were applied to the baseline features. The resulting data form of $X_{trial.rmov}$ obtained from 32 channels of the subject is $32 \times 60 \times 128$, and window segmentation generates the window fragment data X_S^i. The data form of X_S^i is $32 \times 5 \times 128$. Then, X_S^i is divided into five frequency bands, and its data form is expanded to $32 \times 5 \times 128 \times 5$. To extract complete sample timing information and then carry out the time-length 0.5 s framing process, the form is transformed to $32 \times 0.5 \times 128 \times 5$. The five data bands are superimposed in the dimension to convert them into a three-dimensional matrix sequence of size 5×32 with a sequence length of

64 (0.5 × 128) so that each frame enters the convolutional coding module proposed in this paper in the form of 5 × 32 × 64. The important temporal features of EEG bands in matrix sequence data are extracted by Bi-LSTM, and the results of the binary classification on the dataset are shown in Figure 11 below:

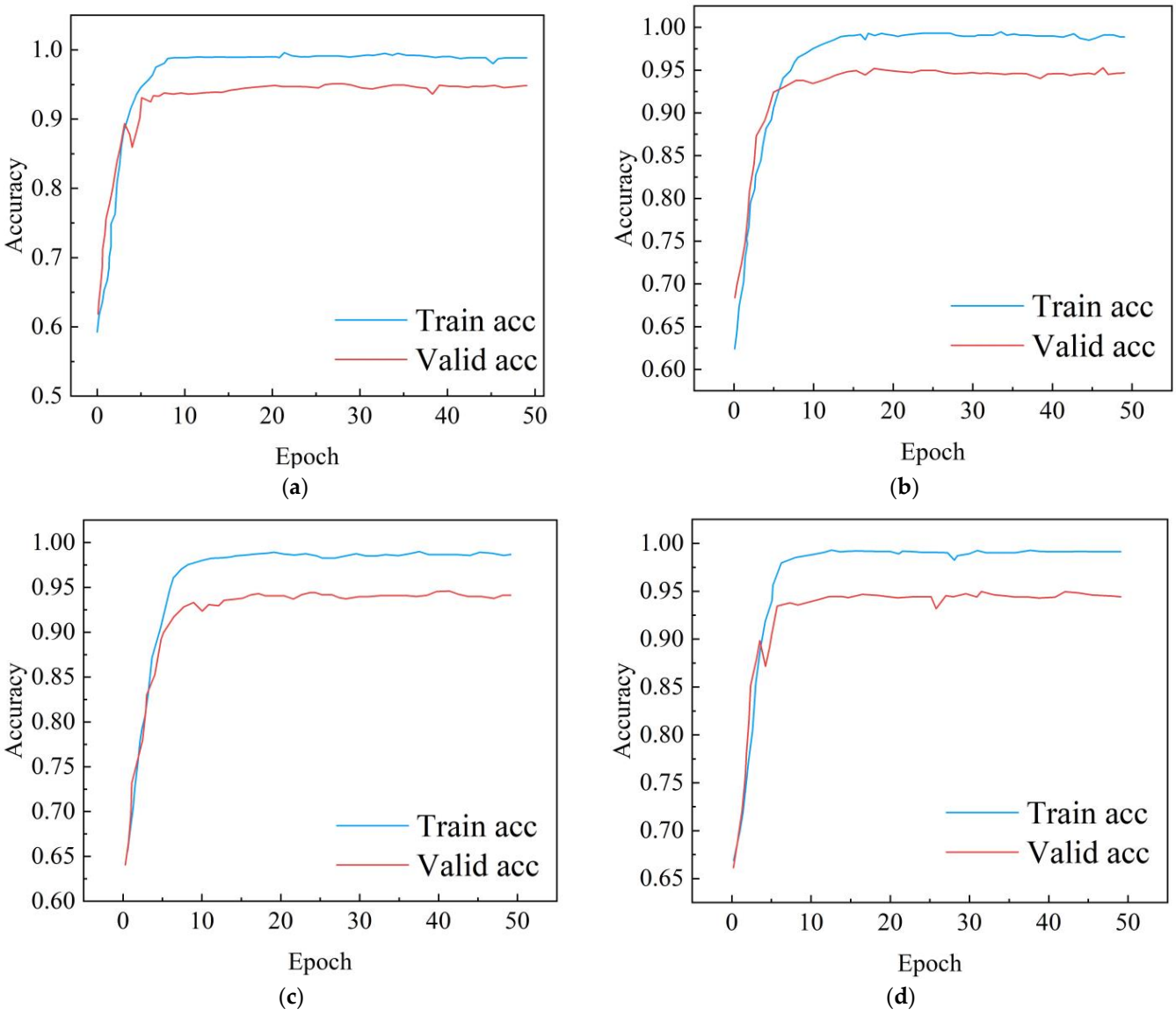


Figure 11. (**a**) Valence dimension recognition results; (**b**) arousal dimension recognition results; (**c**) dominance dimension recognition results; and (**d**) liking dimension recognition results.

The average training and testing accuracies of binary classification for the four evaluation metrics were 98.99% and 94.43%, respectively.

To enhance the spatial representation of the EEG signal, this paper employs a detailed transformation technique that maps the frequency band's sequential signal into a two-dimensional matrix. This approach rectifies the loss of spatial information in the baseline features. Additionally, the more indicative frequency features, namely, DE and PSD, are utilized as the matrix element values to substitute for the sub-band amplitudes in the baseline features. The results of the binary classification on the four DEAP dataset metrics are displayed in Figure 12 below.

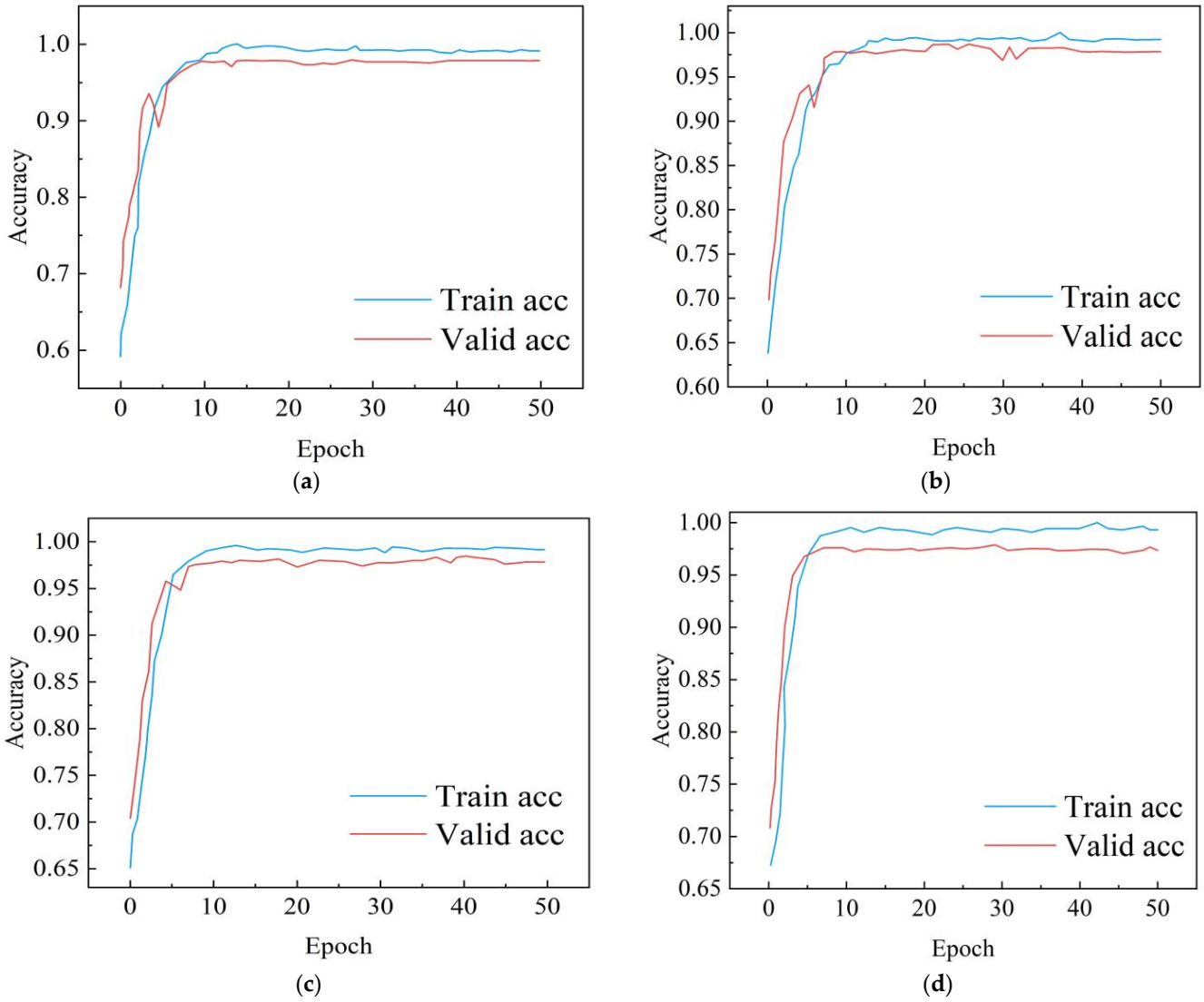


Figure 12. (**a**) Valence dimension recognition results; (**b**) arousal dimension recognition results; (**c**) dominance dimension recognition results; and (**d**) liking dimension recognition results.

It is evident that the average accuracy of the suggested feature extraction method in this article is 99.21% and 97.84% for training and testing in valence, arousal, dominance, and liking states, respectively. This is approximately a 3% improvement over the test accuracy of baseline features.

4.4. Experiment on Four Classes of Emotion Recognition

There were 32 subjects, 40 groups of affective experiments per subject, and 1280 data groups in total. They were labelled and categorized according to four modalities: LALV, LAHV, HALV, and HAHV. Among them, the baseline features have 88.92% and 84.70% training and testing accuracy on the four classifications, respectively, and the feature extraction approach proposed in this article has 90.15% and 88.46% training and testing accuracy, respectively, on the model. The test accuracy improved by nearly 4% compared to the baseline features. The results are shown in Figure 13.

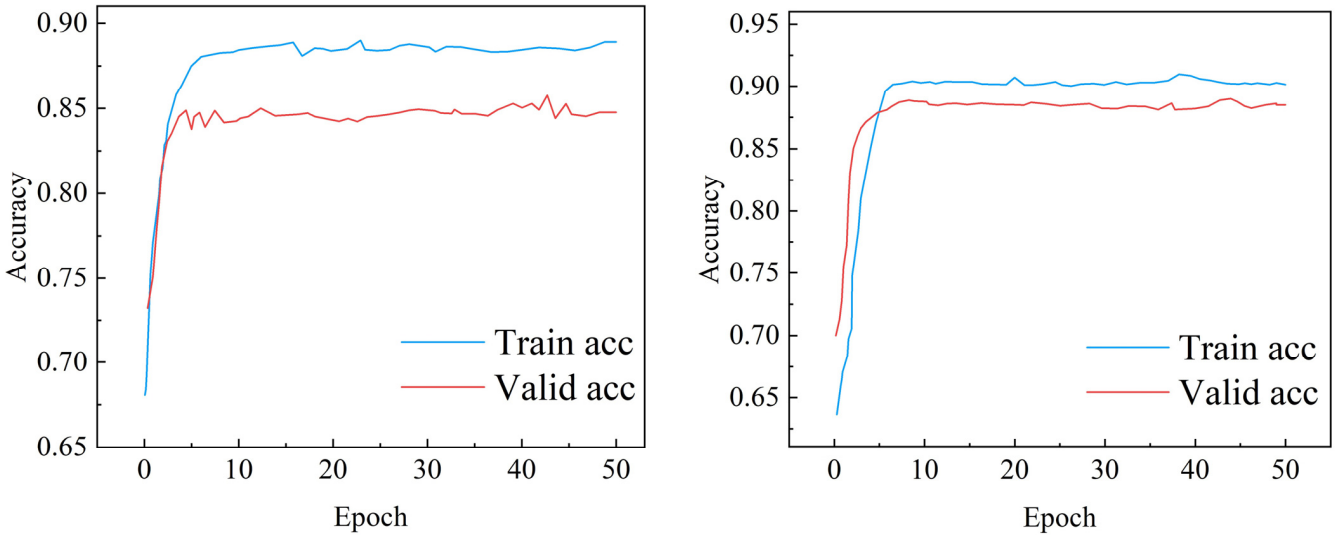


Figure 13. Comparison of valence/arousal four-classification experiments.

4.5. Ablation Experiments

In Section 3.2, three forms of electrode mapping were proposed to determine the optimal mapping method for the deep learning model presented in this article. The performances of these mapping methods were experimentally compared, and the results are presented in Table 4. The sensitive transformation method, which includes more precise electrode locations and provides spatial information that is more consistent with emotional neural features, yielded the highest accuracy. Furthermore, the time consumption of this method was comparable to that of the compact mapping method and approximately one-third of that of the sparse mapping method. Therefore, considering both time and accuracy, the sensitive transformation method is the most suitable approach.

Table 4. Experimental comparison of different mapping methods.

	Shape	**Valence**		**Arousal**		**Valence–Arousal**	
		Acc (%)	**Time (s)**	**Acc (%)**	**Time (s)**	**Acc (%)**	**Time (s)**
Compact Mapping	8 × 9	97.09	89 s	96.85	86 s	87.59	93 s
Sparse Mapping	19 × 19	97.32	255 s	97.06	247 s	88.07	262 s
(Ours)	9 × 9	98.12	95 s	97.72	91 s	88.46	103 s

The present study focuses on the recognition task by employing window segments as samples, which depends on the interframe correlation to extract temporal features. The selection of an appropriate time window and frame length is crucial, and thus, this study conducts experiments to determine optimal values. The accuracy of the model is compared for different values of u and t, as presented in Table 5.

Table 5. Experimental comparison of different time windows and frame lengths.

	DEAP–Valence			**DEAP–Arousal**			**Valence–Arousal**		
	t = 0.25 s	t = 0.5 s	t = 1 s	t = 0.25 s	t = 0.5 s	t = 1 s	t = 0.25 s	t = 0.5 s	t = 1 s
u = 3 s	95.87%	96.62%	94.08%	95.85%	96.23%	93.83%	87.98%	88.11%	85.38%
u = 4 s	96.83%	97.53%	95.02%	97.15%	97.64%	94.58%	88.12%	88.20%	85.05%
u = 5 s	97.43%	98.12%	95.90%	97.30%	97.72%	95.82%	88.30%	88.46%	85.46%
u = 6 s	95.82%	96.72%	95.17%	96.57%	96.89%	94.53%	87.95%	88.14%	85.49%

Based on the results, it is apparent that the optimal classification performance for the DEAP dataset was achieved when u = 5 s and t = 0.5 s. Regarding binary valence and arousal, the maximum difference in accuracy was 4.04% and 3.89%, respectively, and the maximum difference in the four classifications was 3.41%. For window segments of different lengths, the increase in the number of frames enables Bi-LSTM to extract richer temporal information. t = 0.25 s/0.5 s will be higher than the accuracy of 1 s frame division for the same window segment. However, when t = 0.25 s, the doubling of the number of frame sequences does not bring a continuous increase in accuracy, and the overall accuracy is close to the experimental data for t = 0.5 s, but the computational effort of the network is greatly increased.

Furthermore, although the proposed model approach in this paper showed good performance for both binary and four classifications, the effectiveness of individual modules within the model remains unknown. To address this issue, we designed five models with different structures for ablation experiments, and their compositions and results are presented in Table 6. The task identification of all models is based on 4D features designed in this paper, and the baseline CNN is designed without an attention mechanism. The unspecifically labelled LSTM model takes the output of the last time step as the input to softmax. For the binary classification task, the combined CNN and LSTM model achieves an average accuracy of 84.45%, indicating that the combined model can integrate the temporal–frequency–spatial information of the multidimensional features well and produce good results. Bi-LSTM overcomes the limitation of LSTM in learning sequences only sequentially by learning sequences in both directions and combining the hidden layer states to determine the output results, which improves accuracy by approximately 4% compared to the CNN-LSTM model. SENet can focus on important channels in the feature matrix and improve model accuracy by assigning different weights. FcaNet uses two-dimensional discrete cosine variations to compress the feature map, avoiding the loss of frequency components caused by SENet and improving accuracy by 2% at a subtle computational cost. Additionally, to highlight the importance of different frame moments in the sample, the weighted sum of all time step hidden layer states is chosen as the output, achieving the best accuracy of 94.58%.

Table 6. Ablation experiments for different structural models.

Methods	Metrics	Accuracy (%) (avg)	Training Time (s)	Testing Time (s)
CNN-LSTM	Valence Arousal Dominance Liking	84.45	90.51	11.67
CNN-BiLSTM	Valence Arousal Dominance Liking	88.53	93.05	12.26
CNN-SENet-BiLSTM	Valence Arousal Dominance Liking	90.02	86.20	11.47
CNN-FcaNet-BiLSTM	Valence Arousal Dominance Liking	92.25	88.57	11.82
Ours	Valence Arousal Dominance Liking	94.58	89.43	11.87

4.6. Experimental Comparison

In conclusion, the proposed approach is compared with models from other references in the literature, as shown in Tables 7 and 8. In binary classification, some studies, such as [10,11], applied only 1D convolution to extract temporal information in EEG data, whereas Xu [20] constructed 3D spatial–frequency features of DE and selected a fully convolutional residual network recognition model, and Saha [27] used 3D convolutional kernels to extract and process the spatiotemporal features of EEG data. However, none of the above networks fully considered the feature information of the three dimensions, and their average accuracies were lower than the combined model of references [23,24,26], except for the studies where Xu [20] and Saha [27] used channel attention to augment the model accuracy. These findings demonstrate that attention mechanisms and consideration of multidimensional information are significant for the performance improvement of recognition systems. In this article, we employed Bi-LSTM based on LSTM to learn sequence signals for future sentiment information and used FcaNet to compensate for the shortcomings of SENet. Consequently, we achieved the best accuracy of 98.10% for binary classification.

Table 7. Comparison of different deep learning methods for binary classification.

Literature	Metrics	Dataset	Features	Test Accuracy (%)	F1-Score (%)	Methods
Wang Z [13]	Valence Arousal	DEAP	Temporal	83.97 83.72	N/A	Multiscale CNN
Singh K [14]	Valence Arousal H/M/L valence H/M/L arousal	DEAP	Temporal	91.31 (avg) --- 89.32 (avg)	N/A	1DCNN-Bi-LSTM
Bai Z [22]	Valence Arousal Negative/neutral/positive	DEAP --- SEED	Spatial Frequency	88.75 (avg) --- 90.04 (avg)	N/A	CNN (DC + PC + Residual)
Xu X [23]	Negative/neutral/positive	SEED	Spatial Frequency	96.01 (avg)	N/A	CNN (channel attention + Residual)
Meng M [26]	Valence Arousal Negative/neutral/positive	DEAP --- SEED	Frequency temporal spatial	94.85 94.43 --- 94.16 (avg)	N/A	VGG16-LSTM
Li Q [27]	Valence Arousal Negative/neutral/positive	DEAP --- SEED	Frequency temporal spatial	95.02 94.61 --- 95.49 (avg)	96.29 95.72 --- 95.57	CNN-ON-LSTM
Zhang Y [29]	Valence Arousal Negative/neutral/positive	DEAP --- SEED	Frequency temporal spatial	85.86 84.27 --- 92.47 (avg)	N/A	CNN-attention LSTM-attention
Saha O [30]	Valence Arousal	DEAP	Temporal Frequency	97.06 97.34	96.39	3DCNN-channel attention
Ours	Valence Arousal Dominance Liking	DEAP	Frequency temporal spatial	97.84 (avg)	97.22 97.02 97.17 97.11	CNN-BiLSTM-FcaNet

Table 8. Comparison of different deep learning methods for four classifications.

Literature	Metrics	Dataset	Test Accuracy (%)	Methods
Zubair [7]	H/L V-A	DEAP	45.40	mRMR
Guptal [8]	H/L V-A Negative/neutral/positive On beta gamma	DEAP --- SEED	71.43 --- 83.33 (avg)	FAWT-SVM
Aguiñaga [9]	Valence/arousal H/L V-A	DEAP	84.20 (avg) 80.90	WP-NN-SVM
Mei [21]	Valence Arousal H/L V-A Theta alpha gamma	DEAP	83.60 83.0 73.10	CNN
Sharma [39]	Valence Arousal H/L V-A	DEAP	84.16 85.21 82.01	PSO-BiLSTM
Chao [40]	Valence Arousal H/L V-A	DEAP	85.53 85.88 76.77	Multiscale CNN
Ours	H/L V-A	DEAP	88.46	CNN-BiLSTM-FcaNet

In the context of the four-classification task, prior works such as those in [4–6] utilized wavelets to extract time–frequency features for emotion recognition using SVM and mRMR algorithms. However, these methods did not incorporate the spatial information and dynamic temporal features of EEG. Mei [18] and Chao [36] represented features by constructing a connectivity matrix of brain structures, followed by extracting spatial features using CNN. Similarly, PSO-BiLSTM [35] utilized DWT to decompose the signal, applied a third-order cumulant transformation to a high-dimensional space, and then reduced the dimension to eliminate redundancy. However, this approach also did not consider spatial feature learning. In contrast, the four-dimensional data used in this paper contain more information than the two- and three-dimensional data used in prior works. Furthermore, the proposed approach in this article constructs a combined network that can adapt to multidimensional features and extract spatial–frequency and temporal features. It also employs the more advanced FcaNet and fully considers information from all frame moments. These advantages make the proposed method more effective compared to those in prior examples in the literature.

5. Conclusions

This study proposes a cascaded convolutional recurrent neural network based on multidimensional features for emotion recognition. To address the limitations of the previous literature, a 4D matrix is constructed to incorporate emotional features of the signal in the temporal, frequency, and spatial dimensions. Additionally, a hybrid deep learning model is proposed to better fit the extracted feature matrix. The convolutional encoder is mainly used to extract spatial–frequency features from 4D input data, and the residual network composed of DC and PC improves the real-time performance of the recognition system. FcaNet assigns more accurate weights to different feature channels at a negligible computational cost, allowing useful feature information to be further highlighted. Finally, to emphasize the temporal significance of the frame windows in the sample, the weighted sum of the hidden layer states of the Bi-LSTM at all frame moments is utilized as input to the softmax layer. The experimental results demonstrate that the proposed method in this paper performs well compared to the rest of the literature, with an average accuracy of 97.84% in the two classification experiments and 88.46% in the four-classification experiments. In future work, we will explore a more expressive feature extraction method

and apply a more streamlined network to the recognition task, making emotion recognition more rapid in the HCI domain.

Author Contributions: Data curation, C.W.; investigation, Y.L.; resources, Y.L.; software, C.W. and C.L.; supervision, Y.L.; writing—original draft, C.W.; writing—review and editing, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 51775076).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, M.-K.; Kim, M.; Oh, E.; Kim, S.-P. A review on the computational methods for emotional state estimation from the human EEG. *Comput. Math. Methods Med.* **2013**, *2013*, 573734. [CrossRef] [PubMed]
2. Li, Y.; Zheng, W.; Cui, Z.; Zhang, T.; Zong, Y. A Novel Neural Network Model Based on Cerebral Hemispheric Asymmetry for EEG Emotion Recognition. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 1561–1567.
3. Wang, F.; Wu, S.; Zhang, W.; Xu, Z.; Zhang, Y.; Wu, C.; Coleman, S. Emotion recognition with convolutional neural network and EEG-based EFDMs. *Neuropsychologia* **2020**, *146*, 107506. [CrossRef] [PubMed]
4. Bhise, P.R.; Kulkarni, S.B.; Aldhaheri, T.A. Brain computer interface based EEG for emotion recognition system: A systematic review. In Proceedings of the 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 5–7 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 327–334.
5. Prochazka, A.; Vysata, O.; Marik, V. Integrating the role of computational intelligence and digital signal processing in education: Emerging technologies and mathematical tools. *IEEE Signal Process. Mag.* **2021**, *38*, 154–162. [CrossRef]
6. Houssein, E.H.; Hammad, A.; Ali, A.A. Human emotion recognition from EEG-based brain–computer interface using machine learning: A comprehensive review. *Neural Comput. Appl.* **2022**, *34*, 12527–12557. [CrossRef]
7. Zubair, M.; Yoon, C. EEG based classification of human emotions using discrete wavelet transform. In *IT Convergence and Security 2017*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 2, pp. 21–28.
8. Gupta, V.; Chopda, M.D.; Pachori, R.B. Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals. *IEEE Sens. J.* **2018**, *19*, 2266–2274. [CrossRef]
9. Aguiñaga, A.R.; Ramirez, M.A.L. Emotional states recognition, implementing a low computational complexity strategy. *Health Inform. J.* **2018**, *24*, 146–170. [CrossRef]
10. Xing, X.; Li, Z.; Xu, T.; Shu, L.; Hu, B.; Xu, X. SAE+ LSTM: A New framework for emotion recognition from multi-channel EEG. *Front. Neurorobotics* **2019**, *13*, 37. [CrossRef]
11. Ma, J.; Tang, H.; Zheng, W.-L.; Lu, B.-L. Emotion recognition using multimodal residual LSTM network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 176–183.
12. Ye, W.; Li, X.; Zhang, H.; Zhu, Z.; Li, D. Deep Spatio-Temporal Mutual Learning for EEG Emotion Recognition. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–8.
13. Wang, Z. Emotion Recognition Based on Multi-scale Convolutional Neural Network. In Proceedings of the Data Mining and Big Data: 7th International Conference, DMBD 2022, Beijing, China, 21–24 November 2022; Proceedings, Part I. Springer: Berlin/Heidelberg, Germany, 2023; pp. 152–164.
14. Singh, K.; Ahirwal, M.K.; Pandey, M. Quaternary classification of emotions based on electroencephalogram signals using hybrid deep learning model. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 2429–2441. [CrossRef]
15. Catrambone, V.; Greco, A.; Scilingo, E.P.; Valenza, G. Functional linear and nonlinear brain–heart interplay during emotional video elicitation: A maximum information coefficient study. *Entropy* **2019**, *21*, 892. [CrossRef]
16. Zheng, W.-L.; Guo, H.-T.; Lu, B.-L. Revealing critical channels and frequency bands for emotion recognition from EEG with deep belief network. In Proceedings of the 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER), Montpellier, France, 22–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 154–157.
17. Zhang, G.; Yu, M.; Liu, Y.-J.; Zhao, G.; Zhang, D.; Zheng, W. SparseDGCNN: Recognizing emotion from multichannel EEG signals. *IEEE Trans. Affect. Comput.* **2021**, *14*, 537–548. [CrossRef]
18. Hwang, S.; Hong, K.; Son, G.; Byun, H. Learning CNN features from DE features for EEG-based emotion recognition. *Pattern Anal. Appl.* **2020**, *23*, 1323–1335. [CrossRef]

19. Song, T.; Zheng, W.; Song, P.; Cui, Z. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* **2018**, *11*, 532–541. [CrossRef]
20. Cui, G.; Li, X.; Touyama, H. Emotion recognition based on group phase locking value using convolutional neural network. *Sci. Rep.* **2023**, *13*, 3769. [CrossRef] [PubMed]
21. Mei, H.; Xu, X. EEG-based emotion classification using convolutional neural network. In Proceedings of the 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Shenzhen, China, 15–17 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 130–135.
22. Bai, Z.; Liu, J.; Hou, F.; Chen, Y.; Cheng, M.; Mao, Z.; Song, Y.; Gao, Q. Emotion recognition with residual network driven by spatial-frequency characteristics of EEG recorded from hearing-impaired adults in response to video clips. *Comput. Biol. Med.* **2023**, *152*, 106344. [CrossRef]
23. Xu, X.; Cheng, X.; Chen, C.; Fan, H.; Wang, M. Emotion Recognition from Multi-channel EEG via an Attention-Based CNN Model. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, Proceedings of the ICNC-FSKD 2022, Fuzhou, China, 30 July–1 August 2022*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 285–292.
24. Xuan, H.; Liu, J.; Yang, P.; Gu, G.; Cui, D. Emotion Recognition from EEG Using All-Convolution Residual Neural Network. In Proceedings of the Human Brain and Artificial Intelligence: Third International Workshop, HBAI 2022, Held in Conjunction with IJCAI-ECAI 2022, Vienna, Austria, 23 July 2022; Revised Selected Papers. Springer: Berlin/Heidelberg, Germany, 2022; pp. 73–85.
25. Qu, Z.; Zheng, X. EEG Emotion Recognition Based on Temporal and Spatial Features of Sensitive signals. *J. Electr. Comput. Eng.* **2022**, *2022*, 5130184. [CrossRef]
26. Meng, M.; Zhang, Y.; Ma, Y.; Gao, Y.; Kong, W. EEG-based emotion recognition with cascaded convolutional recurrent neural networks. *Pattern Anal. Appl.* **2023**, *26*, 783–795. [CrossRef]
27. Li, Q.; Liu, Y.; Liu, Q.; Zhang, Q.; Yan, F.; Ma, Y.; Zhang, X. Multidimensional Feature in Emotion Recognition Based on Multi-Channel EEG Signals. *Entropy* **2022**, *24*, 1830. [CrossRef] [PubMed]
28. Li, Q.; Liu, Y.; Shang, Y.; Zhang, Q.; Yan, F. Deep Sparse Autoencoder and Recursive Neural Network for EEG Emotion Recognition. *Entropy* **2022**, *24*, 1187. [CrossRef]
29. Zhang, Y.; Zhang, Y.; Wang, S. An attention-based hybrid deep learning model for EEG emotion recognition. *Signal Image Video Process.* **2022**, *17*, 2305–2313. [CrossRef]
30. Saha, O.; Mahmud, M.S.; Fattah, S.A.; Saquib, M. Automatic Emotion Recognition from Multi-Band EEG Data Based on a Deep Learning Scheme with Effective Channel Attention. *IEEE Access* **2022**, *11*, 2342–2350. [CrossRef]
31. Lv, Z.; Zhang, J.; Epota Oma, E. A Novel Method of Emotion Recognition from Multi-Band EEG Topology Maps Based on ERENet. *Appl. Sci.* **2022**, *12*, 10273. [CrossRef]
32. Kamble, K.; Sengupta, J. A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals. *Multimed. Tools Appl.* **2023**, *46*, 1–36. [CrossRef]
33. Frantzidis, C.A.; Lithari, C.D.; Vivas, A.B.; Papadelis, C.L.; Pappas, C.; Bamidis, P.D. Towards emotion aware computing: A study of arousal modulation with multichannel event-related potentials, delta oscillatory activity and skin conductivity responses. In Proceedings of the 2008 8th IEEE International Conference on BioInformatics and BioEngineering, Athens, Greece, 8–10 October 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–6.
34. Li, J.; Zhang, Z.; He, H. Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cogn. Comput.* **2018**, *10*, 368–380. [CrossRef]
35. Shen, F.; Dai, G.; Lin, G.; Zhang, J.; Kong, W.; Zeng, H. EEG-based emotion recognition using 4D convolutional recurrent neural network. *Cogn. Neurodynamics* **2020**, *14*, 815–828. [CrossRef]
36. Xu, W.; Liu, S.; Hou, X.; Yin, X. Sensitive Trans Formation and Multi-Level Spatiotemporal Awareness Based Eeg Emotion Recognition Model. *Adv. Comput. Signals Syst.* **2022**, *6*, 31–41.
37. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
39. Sharma, R.; Pachori, R.B.; Sircar, P. Automated emotion recognition based on higher order statistics and deep learning algorithm. *Biomed. Signal Process. Control* **2020**, *58*, 101867. [CrossRef]
40. Chao, H.; Dong, L. Emotion recognition using three-dimensional feature and convolutional neural network from multichannel EEG signals. *IEEE Sens. J.* **2020**, *21*, 2024–2034. [CrossRef]