mathematics

MDPI

Article

Multi-Channel EEG Emotion Recognition Based on Parallel Transformer and 3D-Convolutional Neural Network

Jie Sun *, Xuan Wang, Kun Zhao, Siyuan Hao and Tianyu Wang

School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266033, China
* Correspondence: sunjie1979@qut.edu.cn

Abstract: Due to its covert and real-time properties, electroencephalography (EEG) has long been the medium of choice for emotion identification research. Currently, EEG-based emotion recognition focuses on exploiting temporal, spatial, and spatiotemporal EEG data for emotion recognition. Due to the lack of consideration of both spatial and temporal aspects of EEG data, the accuracy of EEG emotion detection algorithms employing solely spatial or temporal variables is low. In addition, approaches that use spatiotemporal properties of EEG for emotion recognition take temporal and spatial characteristics of EEG into account; however, these methods extract temporal and spatial information directly from EEG data. Since there is no reconstruction of the EEG data format, the temporal and spatial properties of the EEG data cannot be extracted efficiently. To address the aforementioned issues, this research proposes a multi-channel EEG emotion identification model based on the parallel transformer and three-dimensional convolutional neural networks (3D-CNN). First, parallel channel EEG data and position reconstruction EEG sequence data are created separately. The temporal and spatial characteristics of EEG are then retrieved using transformer and 3D-CNN models. Finally, the features of the two parallel modules are combined to form the final features for emotion recognition. On the DEAP, Dreamer, and SEED databases, the technique achieved greater accuracy in emotion recognition than other methods. It demonstrates the efficiency of the strategy described in this paper.

Keywords: EEG; transformer; 3D-CNN; feature fusion

MSC: 92B20

Citation: Sun, J.; Wang, X.; Zhao, K.; Hao, S.; Wang, T. Multi-Channel EEG Emotion Recognition Based on Parallel Transformer and 3D-Convolutional Neural Network. *Mathematics* **2022**, *10*, 3131. https://doi.org/10.3390/math10173131

Academic Editor: Pedro A. Castillo Valdivieso

Received: 16 July 2022
Accepted: 19 August 2022
Published: 1 September 2022

1. Introduction

Emotions are the attitudes and related behavioral responses of individuals towards objective things. It is a complicated psychological and physiological state resulting from the interaction of emotions, thoughts, and actions. Changes in emotions frequently cause alterations in physiological and non-physiological signals, including intonation [1,2], body posture [2,3], facial expressions [4,5], and the EEG [4]. However, it is difficult to guarantee the reliability and validity of non-physiological signs such as facial expressions, gestures, and speech signals in practical applications [5]. Studies [6] in neuroscience and psychology indicate that EEG can intuitively represent emotional changes in individuals. Furthermore, because EEG signals are difficult to conceal and have outstanding real-time performance, the accuracy is superior to that of other methods, making it one of the most widely used for emotion recognition.

Deep learning is often used for EEG emotion identification since it provides automated feature extraction and classification from beginning to end. To collect EEG data with temporal and spatial characteristics, multiple electrode locations are used. EEG emotion recognition based on deep learning can be loosely divided into temporal recognition methods, spatial recognition methods, and spatiotemporal recognition approaches. Emotion recognition algorithms based on temporal characteristics primarily examine the contextual

characteristics of EEG. Early studies on temporal features were based on manually extracted features [7–11], and deep networks were only used as classifiers. With the development of deep learning, the convolutional neural network (CNN) has been used for temporal feature extraction and has achieved excellent performance in emotion recognition [12–17]. However, CNN often depends on the size of the convolution kernel [18] and has certain defects in extracting global temporal features. The recurrent neural network (RNN) can solve the problem [19], and the long short-term memory model (LSTM) [20] and the gated recurrent network (GRU) [21] are used for the extraction of temporal features and gain satisfactory effects. These approaches are generally effective in extracting temporal features from EEG, but due to the limited amount of data, they are not very reliable at learning the electrode location connection. Methods for emotion recognition based on EEG spatial features account for the spatial interaction between electrodes and rebuild the EEG using electrode spatial information. Early, the spatial features extraction often used the CNN [22–26]. Later, the hybrid networks of CNN and other networks [27,28] perform well in extracting more features from reconstructed images. Such methods take the spatial information of each electrode into account, yet the reconstruction of EEG results in poor contextual feature extraction precision. Therefore, methods based on spatiotemporal characteristics have been proposed to address the abovementioned issues [29–32]. In these methods, temporal features are extracted first, and the features containing temporal information are inputted into the deep network to extract spatial features. However, such algorithms typically utilize only a single type of EEG data to extract features and cannot obtain spatial and temporal details.

To address the aforementioned issues, the study offers an EEG emotion identification model based on the parallel transformer and 3D-CNN spatiotemporal feature fusion. The parallel channel EEG map and position reconstructed EEG sequence are created first. Then, the model transformer [33] and 3D-CNN are employed to extract the respective temporal and spatial information. Finally, the simultaneously extracted spatial and temporal features are connected to a joint feature vector, and the Softmax layer receives it as input to predict emotion.

2. EEG Data Pre-Processing

This section prepares for emotion recognition using EEG. The utilized datasets are detailed in Section 2.1. In Section 2.2, a baseline preprocessing procedure is used to eliminate the impact of individuals' pre-experiment emotions on the results. Section 2.3 creates the parallel channel EEG maps and position reconstructed EEG sequences for the temporal and spatial feature extraction portions.

2.1. EEG Dataset

The DEAP dataset was constructed by Sander Koelstra [34] as a multimodal dataset. The dataset gathered signals from the patients' central and peripheral nerve systems by exposing them to numerous music videos. Utilizing a self-assessment form, the subject's emotional information was gathered. The sample comprised of 32 healthy individuals between the ages of 19 and 37, including 16 males and 16 females. Each subject viewed forty one-minute music videos in accordance with the experimental parameters. According to their feelings, volunteers judged arousal, valence, liking, and dominance after each movie. The score range is 1 to 9. The experimental protocol is depicted in Figure 1. Valence and arousal are the two primary measures utilized to assess human emotions. As valence indicates that people's emotional state goes from negative to positive [35], arousal indicates that people's emotional intensity ranges from weak to robust, with a score over 5 indicating a high emotional state and a score below 5 suggesting a low emotional state.

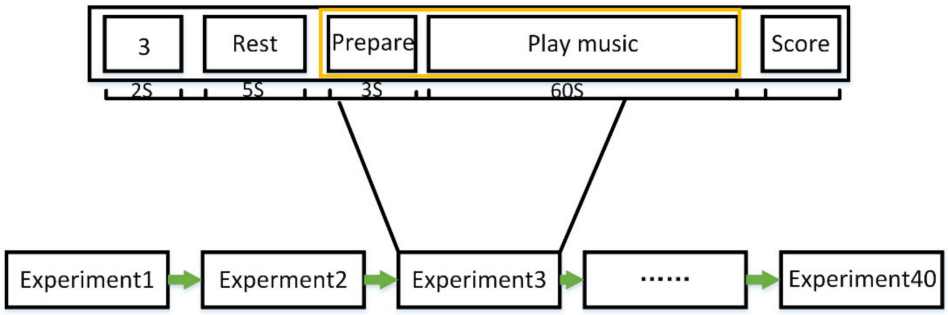


Figure 1. The experimental paradigm for the DEAP dataset. For each experiment, the current experiment number is displayed for 3 s, followed by a 5 s break after the music is played for 63 s, and finally the self-assessment session.

The DEAP dataset contains 40 segments of 32-channel EEG signals from each of 32 participants, with each segment lasting 63 s. The signal has been reduced in frequency to 128 Hz. The initial three seconds of each channel's EEG signal represent the preparatory period, whereas the latter sixty seconds represent the emotion-induced process. We treat the 40-segment EEG signals of a single individual as an independent dataset with the following format: where 40 represents the number of experiments, 32 represents the number of channels, and 8064 represents the number of sampling points.

In addition to the DEAP dataset, the Dreamer and SEED datasets are utilized to validate the model's performance. The Dreamer dataset [36] collected the EEG and ECG signals of 23 subjects during emotion induction; arousal and valence continue to be used as labels for EEG data. If the value is greater than 3, it denotes a high emotional state; otherwise, it suggests a low emotional state. The SEED dataset [37] contains fifteen subjects. Positive, negative, and neutral labels correspond to the target emotion of emotion-evoked segments. The Dreamer and SEED datasets are trained and validated with the same procedure as DEAP.

2.2. Baseline Preprocessing

This article employs the baseline mean approach [38] to determine the respondents' pre-test emotional state. First, remove the pre-test signals from the first three seconds of all channels, divide them into three 1-s segments, and take the mean of the three segments as the baseline mean, which represents the subjects' basic emotions at that moment. Second, the last 60 milliseconds of EEG signal are separated into 60 pieces of equal size, and the baseline effect is subtracted using a Formula (1). Finally, as illustrated in the equation, all baselines shall be spliced (2).

$$Removed_i = Rawdata_i - Basemean \tag{1}$$

$$data = Concat(Removed_1; Removed_2; \ldots; Removed_H) \tag{2}$$

where $Basemean \in R^{C\times L}$, $Rawdata_i \in R^{C\times L}$, $Removed_i \in R^{C\times L}$, $data \in R^{C\times(L\times H)}$; $C = 32$ denotes the number of channels; $L = 1$ s, which represents the length of a segment; $H = 60$ represents the number of segments; *Concat* denotes the connected operations.

Figure 2a depicts the EEG signal of a random channel from the DEAP dataset. Figure 2b depicts the outcome of baseline preprocessing. With baseline interference removed, the EEG signal exhibits a smoother waveform and more pronounced wave features. Following baseline processing, the data format for each participant is where 40 represents the number of experiments, 32 represents the number of channels, and 7680 is the number of sample points in 60 s.

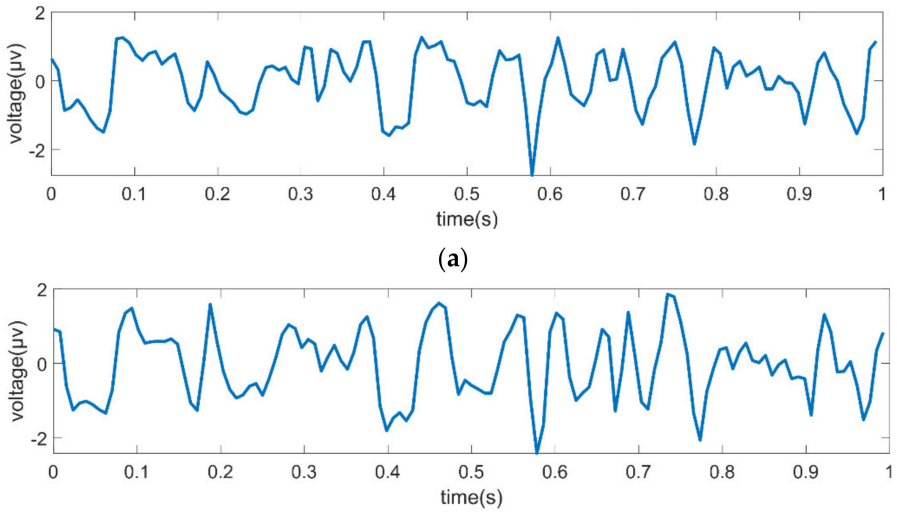


Figure 2. The EEG signal before baseline preprocessing (**a**) and after baseline preprocessing (**b**).

Five people are selected from the DEAP dataset to investigate the impact of baseline preprocessing on emotion recognition performance. Both baseline-preprocessed and unpreprocessed EEG data are used for emotion identification, and the results on arousal and valence are depicted in Figure 3.

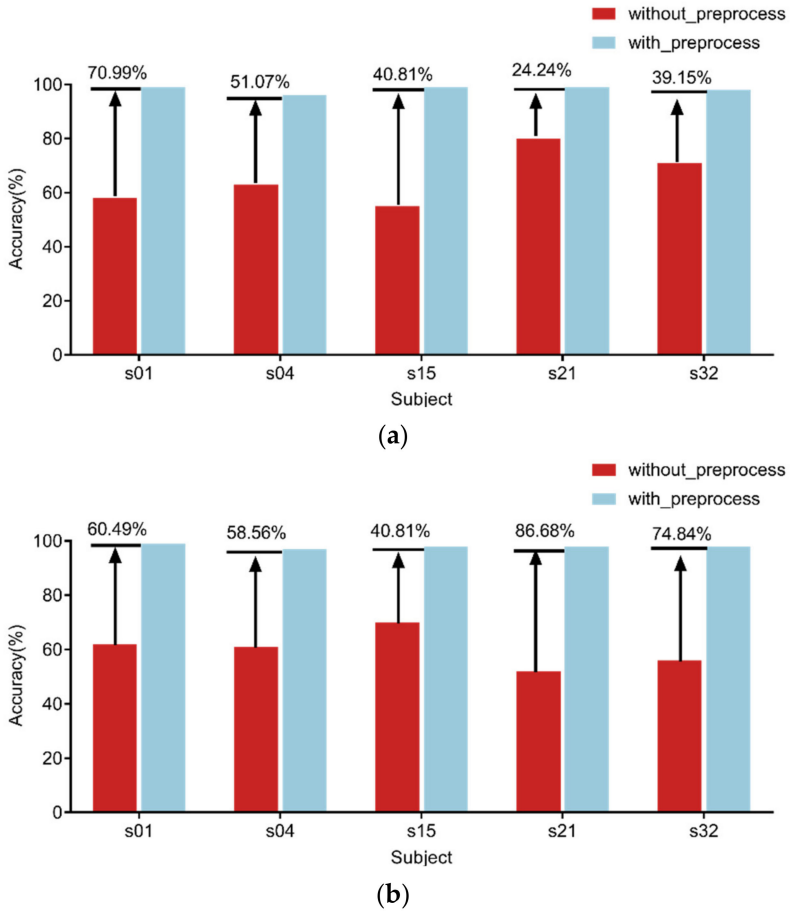


Figure 3. Comparison of accuracy on Arousal (**a**) and Valence (**b**) before and after baseline preprocessing.

Compared to the original EEG, the accuracy of arousal and valence has increased by 24.24 percent and 40.81 percent, respectively, when preprocessed by baseline. Therefore, the baseline is used to preprocess EEG data in this study.

2.3. Construction of Two EEG Representations

To simultaneously gather temporal and spatial features, we need EEG data in two formats that can account for both temporal and spatial aspects. Figure 4 illustrates the construction of parallel channel EEG maps and position-reconstructed EEG sequences.

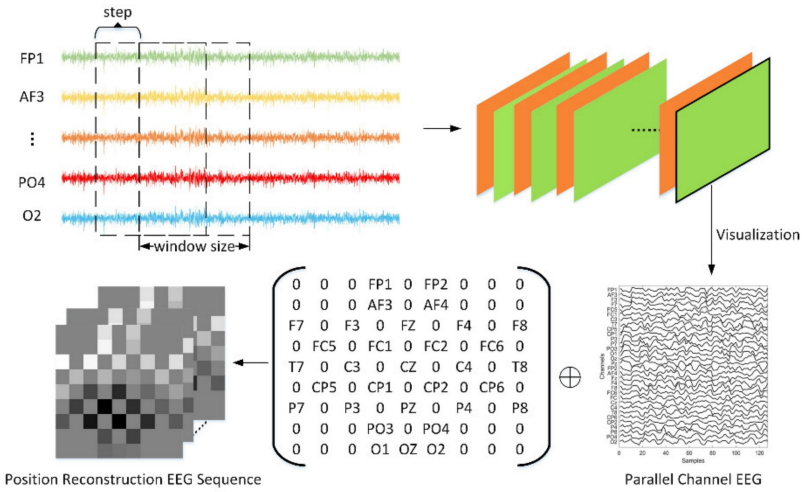


Figure 4. The construction process of two EEG representations. Parallel channel EEG is obtained through a sliding window, and then superimposed with the position projection to convert it to Position Reconstructed EEG sequence.

Initially, the baseline-processed data is divided into EEG segments via a sliding window with all channels placed in a single image, which is referred to as parallel channel EEG. The size 9×9 matrix then stores the parallel channel EEG map based on the electrode distribution map [39]. The placement of the signal of each channel in the corresponding location of the matrix is referred to as the EEG reconstructed sequence. Where 0 indicates the absence of an electrode channel. The data format is then changed from the original matrix to a matrix sequence, which not only retains position information but also maintains temporal characteristics.

To segment EEG data, sliding windows with dimensions of 128, 256, 384, 512 and steps with sizes of 1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8, and 1 of the window widths were selected. Figure 5 depicts the obtained accuracy of arousal and valence. The maximum rate of accuracy is achieved when the window size is 128 and the step size is 1.

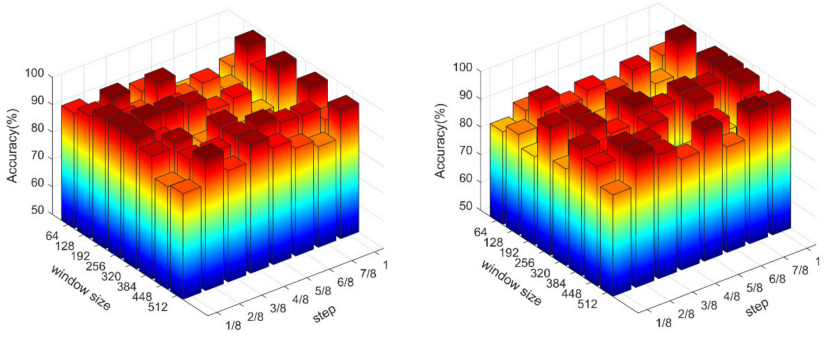


Figure 5. The results on Arousal (**a**) and Valence (**b**) for multiple window size and step parameters.

Each individual can obtain 60 data samples in two formats per experiment, for a total of 2400 EEG samples throughout 40 experiments. A parallel channel EEG map size of 32×128 is utilized to derive temporal characteristics. The sequence size of a position-reconstructed EEG is $9 \times 9 \times 128$, which is used to extract spatial data.

3. Proposed Method

This section proposes TSFFN (Temporal–Spatial Feature Fusion Network), a parallel spatiotemporal feature fusion model based on transformer and 3D-CNN. It includes temporal feature extraction, spatial feature extraction, and temporal-spatial feature fusion modules. In Section 3.1, the attention model transformer in machine translation is utilized to extract the temporal characteristics in order to retrieve the contextual information of the parallel channel EEG maps. In Section 3.2, we employ 3D-CNN to determine the spatial properties of the position-reconstructed EEG episodes. Temporal and spatial features are used as the final emotion detection feature for classification in Section 3.3. The particular procedure is depicted in Figure 6.

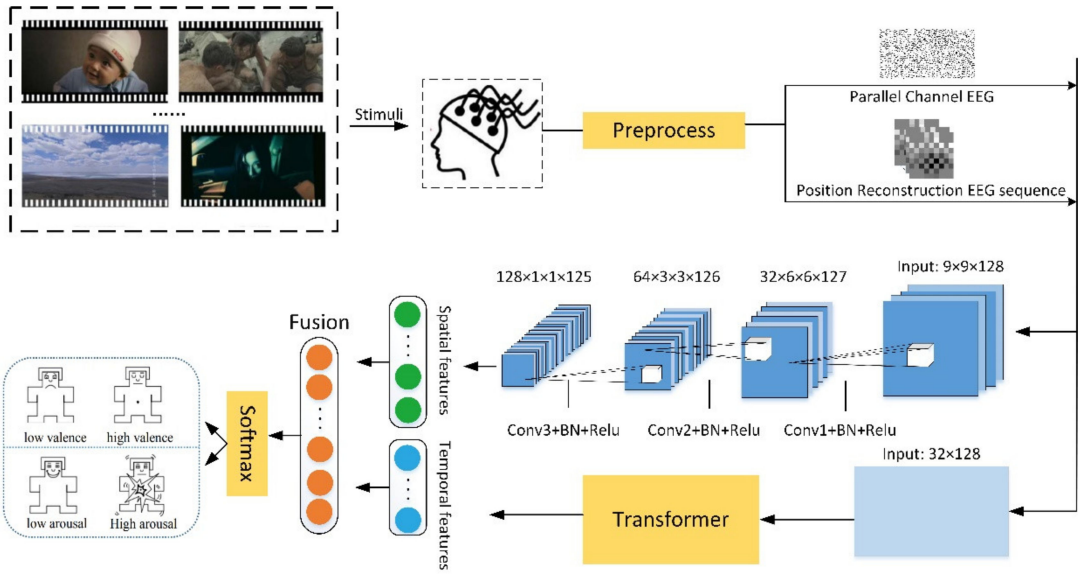


Figure 6. The overall framework of TSFFN. EEG data are simultaneously recorded while subjects watch emotional film clips as stimuli. The spatial and temporal features of EEG are extracted by a parallel model. The emotion predictions are given with the model based on the fused features.

3.1. TSFFN's Temporal Characteristics Extraction Module

This module primarily extracts EEG temporal information, hence EEG parallel channel maps that have been preprocessed serve as input. Considering that the creation of emotions in the brain is a continuous and full process in time, getting the contextual characteristics of EEG is crucial for emotion detection. We used the Transformer to construct a model to extract the temporal characteristics of EEG because the attention mechanism can efficiently perceive the contextual relationships of the signal. The particular procedure is depicted in Figure 7.

Firstly, the position encoding module employs convolution with two channels to obtain the segment position characteristics [40]. The kernel size of convolution influences the accuracy of arousal and valence classification. Table 1 displays the findings of the investigation into the effect of position kernel size on the accuracy of arousal and valence classification. The classification accuracy of arousal and valence is greatest when the convolution kernel size is set at $(1, 51)$. Therefore, the $(1, 51)$ is selected as the position encoding kernel parameter.

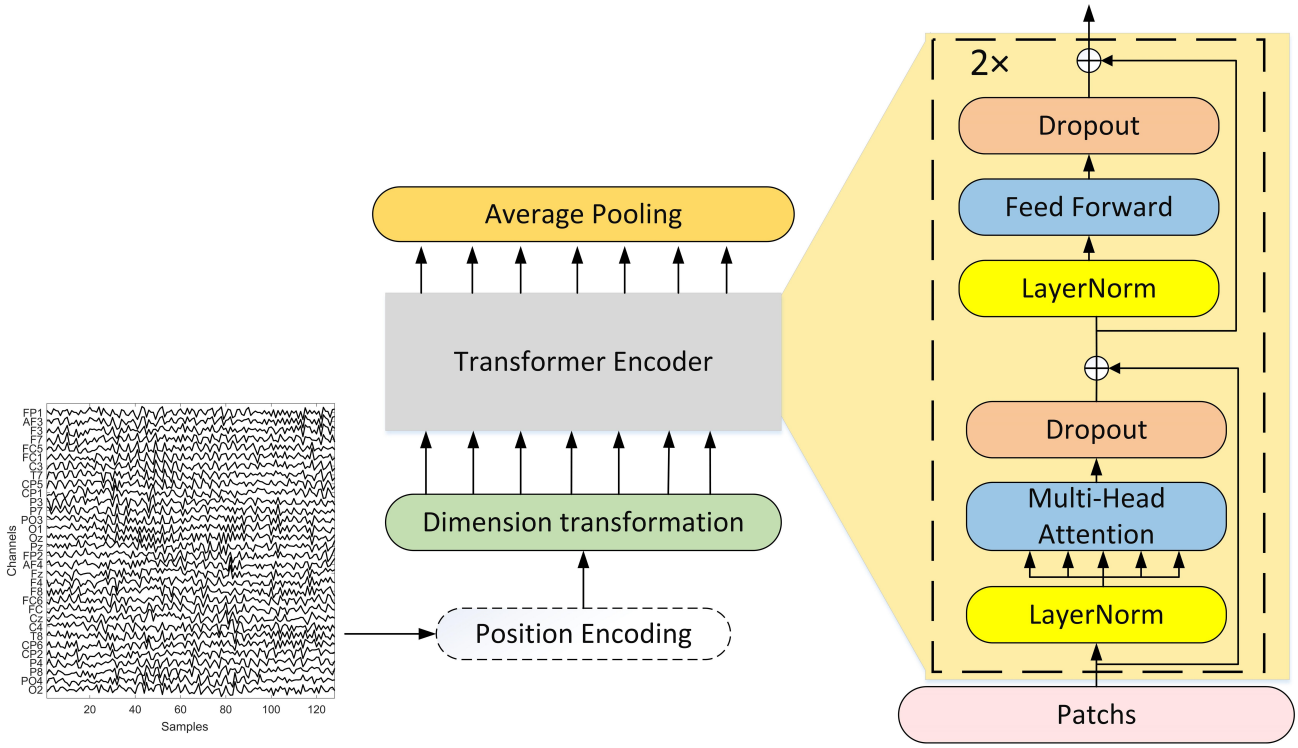


Figure 7. Diagram of temporal feature extraction using the Transformer algorithm.

Table 1. The accuracy of arousal and valence classification based on kernel size.

Kernel Size	Arousal	Valence
(1,1)	96.34%	96.40%
(1,11)	97.16%	97.54%
(1,21)	96.58%	96.50%
(1,31)	97.33%	96.75%
(1,41)	97.29%	97.12%
(1,51)	98.53%	98.27%
(1,61)	97.45%	96.63%
(1,71)	96.83%	96.87%
(1,81)	96.15%	96.08%
(1,91)	97.50%	97.24%

Secondly, the resulting feature maps are then fed into the dimension transformation module, which has a convolution kernel size of (32,16) and a channel count of 10. It transforms the feature maps into seven 1×10 vectors.

Thirdly, the one-dimensional vector sequences are input into the Transformer Encoder, which consists of 2 encoder blocks. After layer normalization, the input is fed into the multi-head attention (MHA) module in the encoder block. The input is turned into five vectors of size 7×2, which are then fed into five heads where the attention mechanism works in parallel. The outputs of all heads are then concatenated and converted to the original size 7×10 following Formulas (3) and (4). At the same time, there is a residual module here.

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^o \tag{3}$$

$$Where\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

where $W_i^Q \in R^{d_{model} \times d_k}$ $W_i^k \in R^{d_{model} \times d_k}$, $W_i^v \in R^{d_{model} \times d_v}$, $W^o \in R^{hd_v \times d_{model}}$ denote the linear transformation used to obtain the query, key and value for each head; $head_i$ represents the outputs of the i-th head after attention mechanism; $Concat$ represents the operation used to concatenate all $head_i$; W^0 indicates a linear change to obtain the final result.

Then the normalization layer's utput is as FF (feed-forward) module's input. The first fully connected layer increases the number of features by four times 7×40. Another full connection layer is connected after the activation function GeLu and the normalization layer to reconstruct the characteristic number to the initial size 7×10.

Finally, a global averaging pooling layer is connected to average the one-dimensional vectors of all channel outputs as the extracted temporal features.

3.2. Spatial Characteristics EXTRACTION Module of TSFFN

The spatial feature extraction module primarily extracts electrode position features, hence the preprocessed position-reconstructed EEG sequence containing electrode position information serves as its input. Since the data comprises both position and temporal information for the electrodes, the 3D convolution [41] is utilized for feature extraction rather than the conventional 2D convolution. The particular procedure is depicted in Figure 8.

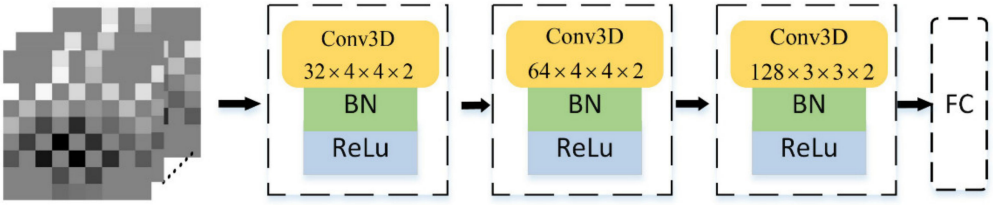


Figure 8. Flow chart of spatial feature extraction based on 3D-CNN. The spatial features of the EEG are extracted through a series of 3D convolution modules and fully connected layers.

The model employs three successive 3D convolutional layers with kernel sizes of 1, 2, and 3, respectively. Convolution kernels in each layer are successively 32, 64, and 128. Because the pooling layer will result in data loss, it is not employed after the convolutional layer to store the electrode position data. Instead of the pooling layer, batch normalization (BN) is introduced to expedite model training. It is then followed by the activation function ReLu. Finally, we apply a complete connection layer to the extracted spatial characteristics to output them. Table 2 displays the model's specific parameters.

Table 2. The structural parameters of 3D-CNN.

Type of Layer		Filter			In Layer	Out Layer	Feature Size
		Width	Height	Length			
conv	Conv1	4	4	2	1	32	$32 \times 6 \times 6 \times 127$
	Conv2	4	4	2	32	64	$64 \times 3 \times 3 \times 126$
	Conv3	3	3	2	64	128	$128 \times 1 \times 1 \times 125$
fc	Fc-512	1	1	125	128	512	512

3.3. Fusion of Temporal and Spatial Modules

Feature fusion and decision fusion are the two most prevalent fusion techniques. The research constructs a TSFFN based on spatiotemporal feature fusion and a TSDFN (Temporal–Spatial Decision Fusion Network) based on spatiotemporal decision fusion to evaluate the performance of the two fusion approaches.

TSFFN concatenates the temporal and spatial features extracted by the transformer and 3D-CNN to obtain a 1×522 vector, which is used as the final feature vector, and then uses a fully connected layer and SoftMax layer for emotion classification. The TSDFN model uses the weighted sum approach [14] to fuse the decision results of the transformer and 3D-CNN, as stated in Formula (5):

$$v_{final} = \sum_{i=1}^{n} w_i v_i \quad (5)$$

Among them, n denotes the number of modules; $v_i = \{+1, -1\}$ is the prediction result of the ith model, where +1 represents a high emotional state, and -1 represents a low emotional state; w_i is the probability of the ith model for emotional prediction; v_{final} is the final decision of the model. If it is greater than or equal to 0, it is judged as a high emotional state, otherwise it is a low emotional state.

In this research, the experimental performance of TSFFN and TSDFN under the two fusion procedures is compared, and the findings are presented in Table 3. Table 3 demonstrates that the classification accuracy of the TSFFN based on feature fusion is higher for both arousal and valence than that of the TSDFN based on decision fusion. Consequently, feature fusion is picked as the model's fusion strategy.

Table 3. The results of different fusion strategies.

Method	Accuracy (Arousal)	Accuracy (Valence)
TSDFN	96.14%	95.76%
TSFFN	98.53%	98.27%

4. Experiments and Analysis of Results

This section verifies the performance of the model. Section 4.1 introduces the hardware and software environment and parameters for model implementation; Section 4.2 describes the model evaluation metrics; Section 4.3 shows the evaluation metrics scores, ablation experiments, and cross-dataset experiments to evaluate the performance of the model; Section 4.4 presents the comparison experiments with other existing EEG emotion recognition algorithms.

4.1. Environment of Software and Hardware for Model Implementation

The proposed model is implemented using Python 3.7 under the PyTorch 1.10.0 framework and trained on the GPU provided by NVIDIA Tesla k80. Python 3.7.0 was released by the python Software Foundation in Delaware, USA, and the python 1.10.0 framework was developed by Facebook in California, USA

4.2. Evaluation Metrics

The Accuracy, Precision, Recall, and F1-score metrics are used to evaluate the performance of the TSFFN. Among them, Accuracy indicates the correct proportion of the model's prediction of all samples; Precision suggests the ratio of the actual positive samples to all the predicted positive samples; Recall indicates the proportion that the correctly predicted positive samples account for all the positive samples; F1-score is a comprehensive indicator of precision and recall, and the calculation formula is as follows:

$$Accurcacy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{8}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \tag{9}$$

where TP represents the prediction is positive and the actual is positive; FP represents that the prediction is positive and the actual is negative; FN represents that the prediction is negative and the actual is positive; TN represents that the prediction is negative and the actual is also negative.

4.3. TSFFN Model Performance Verification

The experiment is carried out on the DEAP dataset. After preprocessing the EEG data of 32 subjects, each subject contains 2400 samples in the form of a parallel channel map and a position reconstructed EEG sequence. The samples are divided into a training set and a test set by 8:2, in which the training set contains 1980 samples, and the test set contains 480 samples. The batch size is 56, epoch is 50, and learning rate is 1×10^{-3}. It minimizes the cross-entropy loss function using the Adam optimizer for training. Moreover, in the case of subject independence, the average training time of the model TSFFN proposed in this paper lasts 714.4523 s.

The TSFFN model is utilized to conduct studies on 32 subjects with the arousal and valence labels, respectively. The results are presented in Tables 4 and 5 respectively.

Table 4. Recognition Accuracy for each subject on "Arousal".

Subject	Accuracy	Subject	Accuracy	Subject	Accuracy	Subject	Accuracy
S01	99.74%	S09	98.77%	S17	97.27%	S25	99.31%
S02	96.74%	S10	99.89%	S18	98.31%	S26	96.2%
S03	99.91%	S11	96.95%	S19	98.37%	S27	98.47%
S04	96.31%	S12	98.66%	S20	99.72%	S28	96.43%
S05	98.22%	S13	99.06%	S21	99.39%	S29	99.33%
S06	98.29%	S14	97.39%	S22	98.29%	S30	98.95%
S07	98.68	S15	99.33%	S23	99.27%	S31	99.12%
S08	99.12	S16	98.85%	S24	99.47%	S32	98.85%

Table 5. Recognition Accuracy for each subject on "Valence".

Subject	Accuracy	Subject	Accuracy	Subject	Accuracy	Subject	Accuracy
S01	99.95%	S09	97.45%	S17	97.62%	S25	97.79%
S02	95.25%	S10	98.87%	S18	98.93%	S26	98.37%
S03	98.97%	S11	95.29%	S19	97.58%	S27	98.97%
S04	97.12%	S12	98.22%	S20	99.25%	S28	97.68%
S05	98.43%	S13	97.47%	S21	98.64%	S29	99.12%
S06	98.89%	S14	97.41%	S22	97.87%	S30	99.14%
S07	99.00%	S15	98.85%	S23	99.62%	S31	98.83%
S08	98.04%	S16	99.29%	S24	98.43%	S32	98.35%

The arousal and valence labeling accuracy of the TSFFN model is greater than 95 percent for all participants, as seen in the tables above. On arousal, the mean $\pm$ standard deviation of the accuracy is 98.53% $\pm$ 1.05% percent, and on valence, it is 98.27% $\pm$ 1.04%.

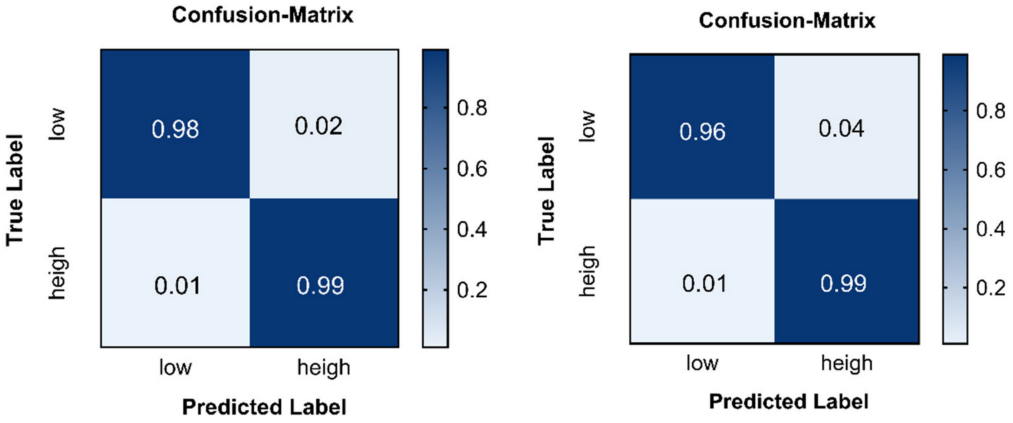
The assessment metrics are computed based on the outcomes of all predictions in the test set, as shown in Table 6. It can be seen that the TSFFN model's arousal scores are higher than its valence ratings. It scored 98.91% on both the Recall and F1-Score tests and 98.92% on the accuracy test. Recall and F1-Score for label valence obtained 98.90% and 98.80% respectively. Precision increased by 98.72% as well.

Table 6. The evaluation metrics scores of TSFFN.

Label	Accuracy	Precision	Recall	F1-Score
Arousal	98.53%	98.92%	98.91%	98.91%
Valence	98.27%	98.72%	98.90%	98.80%

Figure 9 depicts the confusion matrices for the arousal and valence labels. According to the data, the TSFFN has a greater recognition rate for intense emotional states. In the task of label arousal identification, the predictive accuracy of the model for high and low emotions is 99 and 98 percent, respectively. The model also achieves 99 percent

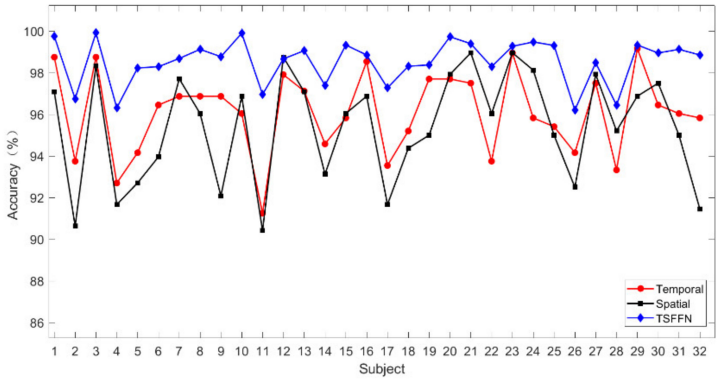
accuracy in valence recognition for high emotional states and 96 percent accuracy for low emotional states.



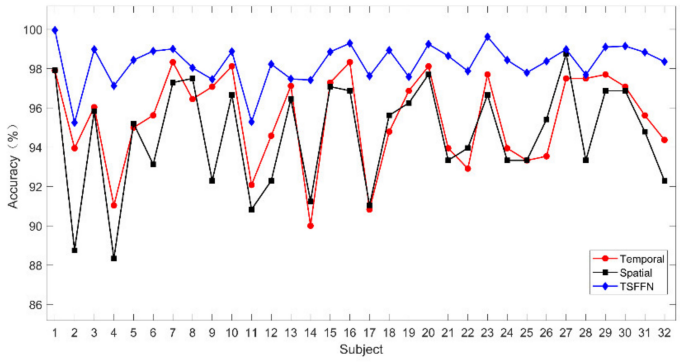
(**a**) (**b**)

Figure 9. Confusion matrices of our method on Arousal (**a**) and Valence (**b**).

The ablation experiments are carried out to verify the validity of each module of the TSFFN. Ablation variables include the temporal feature extraction module (temporal) and the spatial feature extraction module (spatial). Calculate the accuracy of each model's arousal and valence label prediction according to Figure 10.



(**a**)



(**b**)

Figure 10. The results of ablation experiments on Arousal (**a**) and Valence (**b**).

Figure 10 demonstrates that for the two labels of arousal and valence, the accuracy of TSFFN is superior to models based on temporal or spatial variables alone. Therefore, combining spatial and temporal variables can enhance the overall performance of a model.

4.4. TSFFN Performance Testing on Different Datasets

In addition to the DEAP data set, the study also verifies the performance of TSFFN in datasets Dreamer and SEED. On the basis of the results of all subjects, the Accuracy, Precision, Recall, and F1-score indicators are calculated as shown in Table 7, which reveals that on the Dreamer dataset, the accuracy of the model in the arousal label reaches 97.74% and the accuracy in the valence label reaches 96.80%. Additionally, TSFFN scores well in both Precision and Recall. The model also performs exceptionally well on the SEED dataset. The accuracy of the classification of emotions into three categories has reached 97.64%. Therefore, the TSFFFN has good performance not only in DEAP dataset, but also in Dreamer and SEED datasets.

Table 7. The results of cross-dataset experiments.

Index	DEAP		Dreamer		SEED
	Arousal	Valence	Arousal	Valence	
Accuracy	98.53%	98.27%	97.74%	96.80%	97.64%
Precision	98.92%	98.72%	98.66%	98.07%	98.86%
Recall	98.91%	98.90%	98.45%	97.48%	98.86%
F1-score	98.91%	98.80%	98.55%	97.77%	98.86%

4.5. Comparison Experiment with Existing Algorithms

In terms of arousal and valence accuracy, TSFFN compares favorably to PSD [42], DE_CNN [43], IJCNN [38], ACRNN [44], and FSA-3D-CNN [45]. PSD utilized time convolution to obtain spatial features from reconstructed EEG frames and then combines the extracted features with temporal properties for emotion recognition. DE_ CNN described a preprocessing method based on differential entropy (DE) characteristics for transforming raw EEGs to 2D frames, followed by classification of emotions using 3D-CNN. IJCNN captured spatiotemporal characteristics from raw EEG for emotion classification using a hybrid CNN and RNN network. ACRNN provided a hybrid convolutional recurrent network based on attention in which CNN is used to extract spatial features from EEG weighted by the attention mechanism and then an upgraded RNN combines temporal features for emotion recognition; FSA-3D-CNN proposed a four-dimensional representation of the EEG that incorporates spatiotemporal information and then employs a 3D-CNN to classify emotions. Table 8 provides the results. It can be seen from Table 8 that our proposed method has significantly better performance than methods that only use spatial information to reconstruct EEG (DE_CNN, FSA-3D-CNN) or only use time series (ACRNN) as input. In addition, with the method of fusion with spatial and temporal features (PSD, IJCNN), the accuracy of the proposed TSFFN is still far ahead.

Table 8. Performance comparison of different approaches.

Method	Accuracy (Arousal)	Accuracy (Valence)
PSD	85.49%	84.02%
DE_CNN	90.23%	89.78%
IJCNN	91.02%	90.79%
ACRNN	93.81%	92.56%
FSA-3D-CNN	95.87%	95.23%
TSFFN	98.53%	98.27%

5. Limitations

The research has serval limitations, although it achieved greater accuracy in emotion recognition. First, although parallel input can extract more detailed spatial and temporal features, it also requires more runtime. Second, the performance of this study is excellent on independent data sets, but for multi-person diverse data sets, the accuracy of the model will decrease due to individual differences.

In the future, we will try to simplify the network structure and reduce the running time of the model. The model's generalization ability is improved by using multimodal data to minimize the impact of individual differences on the performance of the emotion recognition model.

6. Conclusions

This study introduces TSFFN, a parallel transformer and 3D-CNN-based spatiotemporal feature fusion model. Two forms of EEG are utilized to fully and successfully extract the temporal and spatial characteristics of EEG data. The transformer and 3D-CNN are selected to extract temporal and spatial features from the parallel channel EEG map and reconstructed position EEG sequence, which are then merged to generate the final emotion feature. Not only does the approach simultaneously consider the temporal and spatial aspects of EEG, but it also creates two data forms, which is useful for transformer and 3D-CNN to extract all information. The experimental findings on the DEAP, Dreamer, and SEED datasets demonstrate the adequacy of the suggested model.

Author Contributions: Methodology, J.S. and K.Z.; Software, X.W. and T.W.; Writing—original draft, X.W.; Writing—review & editing, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Nature Fund. Grant number 62171247.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. Song, P.; Jin, Y.; Zhao, L.; Xin, M. Speech Emotion Recognition Using Transfer Learning. *IEICE Trans. Inf. Syst.* **2014**, *97*, 2530–2532. [CrossRef]
2. Yan, J.; Zheng, W.; Xin, M.; Qiu, W. Bimodal emotion recognition based on body gesture and facial expression. *J. Image Graph.* **2013**, *23*, 333–337.
3. Huang, X.; Zhao, G.; Hong, X.; Zheng, W.; Pietikäinen, M. Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns. *Neurocomputing* **2016**, *175 Pt A*, 564–578. [CrossRef]
4. Zheng, W. Multichannel EEG-Based Emotion Recognition via Group Sparse Canonical Correlation Analysis. *IEEE Trans. Cogn. Dev. Syst.* **2017**, *9*, 281–290. [CrossRef]
5. Kim, K.H.; Bang, S.W.; Kim, S.R. Emotion recognition system using short-term monitoring of physiological signals. *Med. Biol. Eng. Comput.* **2004**, *42*, 419–427. [CrossRef] [PubMed]
6. Alarcao, S.M.; Fonseca, M.J. Emotions Recognition Using EEG Signals: A Survey. *IEEE Trans. Affect. Comput.* **2017**, *10*, 374–393. [CrossRef]
7. Zheng, W.L.; Zhu, J.Y.; Lu, B.L. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* **2019**, *10*, 417–429. [CrossRef]
8. Wang, X.H.; Zhang, T.; Xu, X.M.; Chen, L.; Xing, X.F.; Chen, C.P. EEG emotion recognition using dynamical graph convolutional neural networks and broad learning system. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018; IEEE Press: Madrid, Spain, 2018; Volume 1, pp. 1240–1244.
9. Qiao, R.; Qing, C.; Zhang, T.; Xing, X.; Xu, X. A novel deep-learning based framework for multi-subject emotion recognition. In Proceedings of the 2017 4th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS), Dalian, China, 24–26 July 2017; IEEE Press: Dalian, China, 2017; pp. 181–185.
10. Li, Y.; Huang, J.; Zhou, H.; Zhong, N. Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. *Appl. Sci.* **2017**, *7*, 1060. [CrossRef]
11. 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. Neurorobot.* **2019**, *13*, 37. [CrossRef]
12. Lin, C.-T.; Chuang, C.-H.; Hung, Y.-C.; Fang, C.-N.; Wu, D.; Wang, Y.-K. A Driving Performance Forecasting System Based on Brain Dynamic State Analysis Using 4-D Convolutional Neural Networks. *IEEE Trans. Cybern.* **2020**, *51*, 4959–4967. [CrossRef]
13. Amin, S.U.; Alsulaiman, M.; Muhammad, G.; Mekhtiche, M.A.; Hossain, M.S. Deep Learning for EEG motor imagery classification based on multi-layer CNNs feature fusion. *Future Gener. Comput. Syst.* **2019**, *101*, 542–554. [CrossRef]
14. Yang, H.; Han, J.; Min, K. A Multi-Column CNN Model for Emotion Recognition from EEG Signals. *Sensors* **2019**, *19*, 4736. [CrossRef]

15. Wei, C.; Chen, L.L.; Song, Z.Z.; Lou, X.G.; Li, D.D. EEG-based emotion recognition using simple recurrent units network and ensemble learning. *Biomed. Signal Process. Control* **2020**, *58*, 101756. [CrossRef]
16. Lu, Z. An Experimental Study on Relationship Between Subliminal Emotion and Implicit Sequence Learning: Evidence from Eye Movements. *Int. J. Psychol. Brain Sci.* **2018**, *3*, 1. [CrossRef]
17. Li, X.; Song, D.; Zhang, P.; Yu, G.; Hou, Y.; Hu, B. Emotion recognition from multi-channel EEG data through Convolutional Recurrent Neural Network. In Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine, Kansas City, MO, USA, 13–16 November 2017.
18. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral Image Classification Using the Bidirectional Encoder Representation from Transformers. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 165–178. [CrossRef]
19. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Li, Y. Spatial–Temporal Recurrent Neural Network for Emotion Recognition. *IEEE Trans. Cybern.* **2019**, *49*, 839–847. [CrossRef]
20. Jeevan, R.K.; Rao, S.; Kumar, P.S.; Srivikas, M. EEG-based emotion recognition using LSTM-RNN machine learning algorithm. In Proceedings of the 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 25–26 April 2019.
21. Lew, W.C.L.; Wang, D.; Shylouskaya, K.; Zhang, Z.; Lim, J.H.; Ang, K.K.; Tan, A.H. EEG-based Emotion Recognition Using Spatial-Temporal Representation via Bi-GRU. In Proceedings of the IEEE Annual International Conference of the Engineering in Medicine and Biology, Montreal, QC, Canada, 20–24 July 2020; pp. 116–119.
22. Chao, H.; Dong, L.; Liu, Y.; Lu, B. Improved Deep Feature Learning by Synchronization Measurements for Multi-Channel EEG Emotion Recognition. *Hindawi* **2020**, *2020*, 6816502. [CrossRef]
23. Song, T.; Zheng, W.; Song, P.; Cui, Z. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* **2020**, *11*, 532–541. [CrossRef]
24. Li, J.; Zhang, Z.; He, H. Hierarchical convolutional neural networks for eeg-based emotion recognition. *Cogn. Comput.* **2017**, *10*, 368–380. [CrossRef]
25. Robinson, N.; Lee, S.; Guan, C. EEG representation in deep convolutional neural networks for classification of motor imagery. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 1322–1326.
26. Cho, J.; Hwang, H. Spatio-temporal representation of an electoencephalogram for emotion recognition using a three-dimensional convolutional neural network. *Sensors* **2020**, *20*, 3491. [CrossRef]
27. Bagherzadeh, S. *A Hybrid Eeg-Based Emotion Recognition Approach Using Wavelet Convolutional Neural Networks (Wcnn) and Support Vector Machine*; Negah Scientific Publisher: Münster, Germany, 2021.
28. Dai, G.; Zhou, J.; Huang, J.; Wang, N. Hs-cnn: A cnn with hybrid convolution scale for eeg motor imagery classification. *J. Neural Eng.* **2020**, *17*, 016025. [CrossRef] [PubMed]
29. Ding, Y.; Robinson, N.; Zhang, S.; Zeng, Q.; Guan, C. TSception: Capturing Temporal Dynamics and Spatial Asymmetry from EEG for Emotion Recognition. *arXiv* **2021**, arXiv:2104.02935. [CrossRef]
30. Li, X. Motor imagery-based EEG signals classification by combining temporal and spatial deep characteristics. *Int. J. Intell. Comput. Cybern.* **2020**, *13*, 437–453.
31. Qiao, W.; Bi, X. Deep Spatial-Temporal Neural Network for Classification of EEG-Based Motor Imagery. In Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Wuhan, China, 12–13 June 2019.
32. Wang, Z.; Wang, Y.; Zhang, J.; Hu, C.; Yin, Z.; Song, Y. Feature Fusion Based Deep Residual Networks Using Deep and Shallow Learning for EEG-Based Emotion Recognition. *Chin. J. Biomed. Eng.* **2021**, *40*, 641–652.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
34. Koelstra, S. DEAP: A Database for Emotion Analysis; Using Physiological Signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [CrossRef]
35. Russell, J.A. A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
36. Katsigiannis, S.; Ramzan, N. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 98–107. [CrossRef]
37. Duan, R.N.; Zhu, J.Y.; Lu, B.L. Differential entropy feature for EEG-based emotion classification. In Proceedings of the 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), San Diego, CA, USA, 6–8 November 2013.
38. Yang, Y.; Wu, Q.; Qiu, M.; Wang, Y.; Chen, X. Emotion recognition from multichannel EEG through parallel convolutional recurrent neural network. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; IEEE Press: Rio de Janeiro, Brazil, 2018; pp. 1–7.
39. Oostenveld, R.; Praamstra, P. The five percent electrode system for high-resolution EEG and ERP measurements. *Clin. Neurophysiol.* **2001**, *112*, 713–719. [CrossRef]
40. Song, Y.; Jia, X.; Yang, L.; Xie, L. Transformer-based Spatiotemporal Feature Learning for EEG Decoding. *arXiv* **2021**, arXiv:2106.11170.
41. Li, Y.; Li, Y.; Yang, B. A review of EEG emotion recognition based on deep learning. *Beijing Biomed. Eng.* **2020**, *39*, 634–642.
42. Liu, W.; Qiu, J.L.; Zheng, W.L.; Lu, B.L. Emotion Recognition of EEG Signals Based on Location Information Reconstruction and Time-frequency Information Fusion. *Comput. Eng.* **2021**, *47*, 95–102.

43. Yang, Y.; Wu, Q.; Fu, Y.; Chen, X. Continuous Convolutional Neural Network with 3D Input for EEG-Based Emotion Recognition. In Proceedings of the International Conference on Neural Information Processing, Siem Reap, Cambodia, 13–16 December 2018; Springer: Cham, Switzerland, 2018.
44. Tao, W.; Li, C.; Song, R.; Cheng, J.; Liu, Y.; Wan, F.; Chen, X. EEG-based Emotion Recognition via Channel-wise Attention and Self Attention. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
45. Zhang, J.; Zhang, X.; Chen, G.; Yan, C. EEG emotion recognition based on the 3D-CNN and spatial-frequency attention mechanism. *J. Xidian Univ.* **2022**, 1–9.