mathematics

MDPI

Article

A Comprehensive Interaction in Multiscale Multichannel EEG Signals for Emotion Recognition

Yiquan Guo, Bowen Zhang, Xiaomao Fan, Xiaole Shen and Xiaojiang Peng *

College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China; 2110416011@stumail.sztu.edu.cn (Y.G.); zhangbowen@sztu.edu.cn (B.Z.); fanxiaomao@sztu.edu.cn (X.F.); shenxiaole@sztu.edu.cn (X.S.)

* Correspondence: pengxiaojiang@sztu.edu.cn

Abstract: Electroencephalogram (EEG) is the most preferred and credible source for emotion recognition, where long-short range features and a multichannel relationship are crucial for performance because numerous physiological components function at various time scales and on different channels. We propose a cascade scale-aware adaptive graph convolutional network and cross-EEG transformer (SAG-CET) to explore the comprehensive interaction between multiscale and multichannel EEG signals with two novel ideas. First, to model the relationship of multichannel EEG signals and enhance signal representation ability, the multiscale EEG signals are fed into a scale-aware adaptive graph convolutional network (SAG) before the CET model. Second, the cross-EEG transformer (CET), is used to explicitly capture multiscale features as well as their correlations. The CET consists of two self-attention encoders for gathering features from long-short time series and a cross-attention module to integrate multiscale class tokens. Our experiments show that CET significantly outperforms a vanilla unitary transformer, and the SAG module brings visible gains. Our methods also outperform state-of-the-art methods in subject-dependent tasks with 98.89%/98.92% in accuracy for valence/arousal on DEAP and 99.08%/99.21% on DREAMER.

Keywords: EEG classification; emotion recognition; multiscale feature; cross attention

MSC: 92B20

Citation: Guo, Y.; Zhang, B.; Fan, X.; Shen, X.; Peng, X. A Comprehensive Interaction in Multiscale Multichannel EEG Signals for Emotion Recognition. *Mathematics* **2024**, *12*, 1180. https://doi.org/10.3390/math12081180

Academic Editor: Jonathan Blackledge

Received: 9 March 2024
Revised: 8 April 2024
Accepted: 9 April 2024
Published: 15 April 2024

1. Introduction

Affective computing is an umbrella term for human emotion, sentiment, and emotion recognition. As emotion affects human daily behaviors and cognitive activities, emotion recognition plays a crucial role in research fields such as artificial intelligence, medical and health [1–6], and brain–computer interfaces (BCI). Generally, both nonphysiological and physiological signals can be used to recognize human emotion [7]. The study conducted by Cacioppo et al. [8] indicates that changes in emotions can lead to physiological variations, including facial muscle movements, brain activity (EEG), and autonomic nervous system (ANS) activity. In brain activity, highly excited or anxious emotional states result in higher-frequency, more intense, and more irregular electrical activity in the brain. Conversely, relaxed and calm emotional states may manifest as lower-frequency, weaker, and more regular electrical activity. Compared to nonphysiological signals such as facial expression, body behavior, speech signals, and textual information, physiological signals have the advantage of being difficult to fake [9], making them more credible for emotion recognition. Among physiological signals, the EEG signal directly reflects brain activity during an emotional response, which is proved to be more effective than other noninvasive physiological signals, such as electrocardiogram (ECG), electrooculogram (EOG), galvanic skin response (GSR), electromyogram (EMG), humidity, and temperature.

In the early years, most of the works decoupled the EEG-based emotion recognition as a feature extraction stage and a classification stage. Feature extraction is the vital stage

of emotion recognition, which can be conducted on both the temporal domain and the spectral domain. Temporal domain features mainly reflect the temporal information of EEG signals, where typical ones include sample entropy [10] and fractal dimension feature [11]. Spectral domain features aim to capture emotion information in different frequency bands, where commonly used ones include differential entropy (DE) [12], power spectral density (PSD) [13], etc. These features mainly rely on researchers' careful design with professional knowledge, which can be biased and limited in representative ability.

Recently, with the great success of deep learning methods in most of the research fields [14–16], deep learning-based EEG emotion recognition methods, which can unify the feature extraction and classification stages in an end-to-end manner, emerged in large numbers. Regarding the multiple channels of EEG signals as spatial information, a straightforward scheme is to apply 2D convolutional neural networks (CNNs) to these spatial–temporal series [17,18]. Generally, a recurrent neural network (RNN) or long short-term memory network (LSTM) can be further used to gather temporal information on CNN feature maps [19–22]. Optionally, transformers [6,23] or 3D CNNs can be also applied if EEG signals are viewed as temporal segments [24]. As the EEG signals are discontinuous in the spatial domain, it is suggested that it may be not suitable to apply CNNs in this way on images. More recently, there is a trend to build models based on graph theory, where the graph is mainly designed or learnable, to describe the relation of multiple channels [25–29].

Among the aforementioned methods, almost all of them do not pay attention to multiscale EEG information. However, the physical sciences have indicated that numerous physiological components function at various time scales [30,31], suggesting the importance of multiscale information for EEG signals. To this end, there exist some other studies which focus on multiscale EEG features-based emotion recognition, where multiscale permutation entropy (MPE) and multiscale convolutional kernels are commonly used [32–34]. Overall, there is rare exploration on the views of both multiscale information and spatial relationships.

In this paper, we jointly consider the multiscale information and the multichannel relationship of raw EEG signals and propose a conceptually new yet simple method, termed the cascade scale-aware adaptive graph and cross-EEG transformer (SAG-CET). Unlike the recent transformer-based EEG emotion recognition methods that make efforts on single-transformer architecture [6,23], our SAG-CET builds a scale-aware adaptive GCN and a dual-stream transformer to integrate both the interchannel relation and multiscale information. Specifically, EEG signals with multiple channels are first split into nonoverlapping multichannel samples, as is common, and then we apply two division schemes on samples to generate short-scale and long-scale patches. After linear projection, both kinds of patches are fed into the SAG to enhance their representative abilities from the relation of different EEG channels. Subsequently, these patches along with two class tokens are fed into the CET. In the CET, different scale patches and tokens are first encoded by separated self-attention (SA) encoders, and then the class tokens are exchanged twice in a cross-attention operation to gather interscale correlation information. We apply several CETs to enhance model capacity and performance and finally take both class tokens for classification.

Additionally, to deeply investigate the effects of different scale information, we conduct extensive evaluations on DEAP and DREAMER to answer the following questions: (1) How does signal scale affect emotion recognition? (2) How does spatial correlation affect emotion recognition? (3) What are the optimal scales for SAG-CET? We observe that our SAG-CET significantly outperforms unitary transformer-based EEG emotion recognition yet is better than feature concatenation strategy and late fusion. We finally set up new state-of-the-art performance on two popular EEG emotion datasets. On DEAP, we achieve 98.89% and 98.92% in average subject-dependent accuracy of valence and arousal, and the numbers are 99.03% and 99.21% on DREAMER.

Our contributions can be summarized as follows:

- We propose to jointly consider the spatial relationship and temporal scale for EEG-based emotion recognition with a novel cascade framework.

- With multiscale EEG patch signals, we introduce the SAG to learn the relationship of multiple channels and propose the CET, which is superior to other fusion strategies, to better integrate multiscale representations.
- We conduct extensive experiments on the issues of spatial relationship and temporal scale and set up new state-of-the-art performance on two popular EEG emotion datasets.

2. Related Work

2.1. Valence–Arousal Model

The dimensional theory proposed by Russell et al. [35] categorizes emotions based on dimensional space. This theory suggests that emotions are constantly changing. One approach represents emotions using two variables, valence and arousal. In most emotion recognition studies, different emotions are mainly classified according to the valence–arousal emotional model. As shown in Figure 1, the vertical axis signifies arousal, which describes the progression of emotions from calm to excitement. The horizontal axis represents valence, reflecting emotions from low to high levels of positivity. Different emotions can be indicated by various coordinates on the graph; for example, excitement corresponds to positive arousal and positive valence, while sadness corresponds to negative arousal and negative valence.

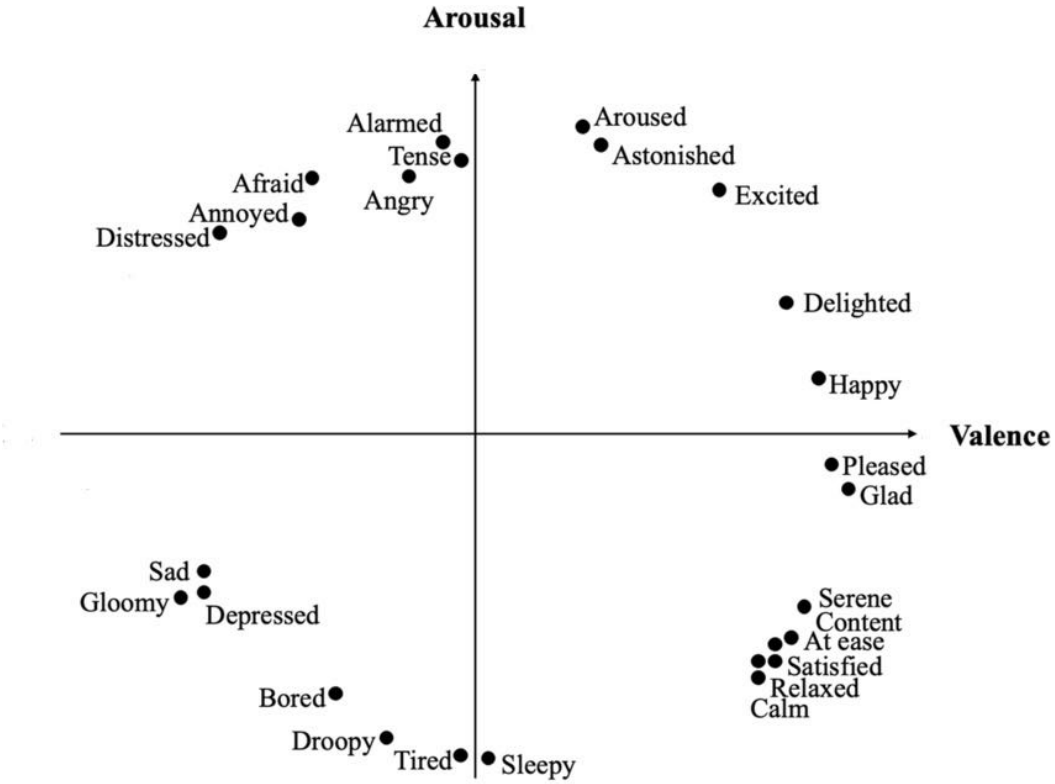

Figure 1. Valence–arousal model.

2.2. EEG Emotion Recognition

The initial emotion recognition is mainly dominated by traditional machine learning. Li et al. [36] selected the frequency band in the spectral domain to extract relevant features and used a common spatial pattern and support vector machine (SVM) for classification. Bazgir et al. [37] used wavelet transform and principal component analysis to extract the frequency domain features of EEG signals and, respectively, used machine learning methods such as k-nearest neighbor (KNN) SVM to classify emotions. Smith K. Khare et al. [38] proposed optimized variational mode decomposition for emotion recognition using single-channel EEG signals. These methods based on traditional machine learning

require good feature design methods, and as the amount of data increases, the performance of the model is affected.

The subsequently developed deep learning algorithms performed well despite the increasing data. For example, Zhong et al. [39] proposed a regionally related and attention-driven bidirectional LSTM network (RA-BiLSTM) for classifying brain activity induced by images. Li et al. [40] designed a hybrid deep learning model that combines CNN and RNN to extract features. These methods achieved good results but encountered new problems: Existing deep learning methods cannot focus on the important characteristic of long-range correlation in the temporal scale of EEG signals. To solve the aforementioned issue, researchers have attempted to use transformers. Li et al. [6] proposed an automatic transformer neural architectures search (TNAS) framework based on a multiobjective evolution algorithm (MOEA) for EEG-based emotion recognition.

2.3. The Usage of Multichannel and Multiscale Information

The spatial distribution of EEG electrodes is adjacent. Due to the limitation of insufficient spatial resolution of EEG electrodes, the collected EEG signal of one electrode will be affected by the EEG signals from other electrodes. Therefore, it is necessary to model the influence of different spatial regions using spatial relationships [41]. To this end, some researchers process EEG signals in a way similar to the field of computer vision to enable the model to automatically capture spatial features by convolution operation, as shown in the middle of Figure 2. For example, Tao et al. [42] proposed ACRNN, which maps the EEG signals onto a matrix similar to an image to represent the spatial relationships between multiple channels. However, the problem is that the interpolation of 2D matrix mapping introduces noise at discontinuous points between different EEG channels and cannot effectively represent the relationship between multiple channels. Some researchers resorted to graph theory to construct the relationship of multiple channels, as shown in the right of Figure 2. For example, Jia et al. [43] proposed a novel heterogeneous graph recurrent neural network (HetEmotionNet), fusing multimodal physiological signals for emotion recognition.

Figure 2. Previous usages of multichannel EEG signals. Left: the spatial distribution of EEG electrodes. Middle: CNN for spatial relationship feature extraction and recognition. The disadvantage of this approach is that there are a large number of interpolated 0 values as noise that interfere with the spatial relationship features. Right: modeling spatial relation as graph.

In addition to using graph theory to model the relations of multiple channels, the utilization of multiscale features also greatly improves the performance [30,31]. Jomaa et al. [44] introduced an MPE-based method that measures complexity in non-stationary multivariate signals. Su et al. [45] proposed a 3D CNN model with multiscale convolutional kernels to recognize emotional states. Although the aforementioned methods attempted to utilize the multiscale features of EEG signals and achieved satisfactory performance, features like MPE and multiscale convolution kernels still need predesign and computation. In addition, few studies combine multiscale features with multichannel features constructed from a graph.

Overall, although existing EEG emotion recognition methods have achieved good performance, there is still a lack of comprehensive methods to address the aforementioned issues. Our proposed method, SAG-CET, effectively utilizes a combination of graph theory and a transformer. Additionally, SAG-CET aims to streamline and improve the representation capabilities of EEG signals by capturing inherent time multiscale features.

3. Methodology

3.1. Preliminaries

Vision transformer: A vision transformer (ViT) is a variant architecture of the transformer [16] used in computer vision [14]. Different from the transformer that deals with 1D data in the field of NLP, the VIT applied in an image handles 2D data. The ViT is composed of multiple multihead attention modules and multiple dense layers, which are combined alternately.

The ViT divides a 2D image $x \in \mathbb{R}^{H\times W\times C}$ into N patches $x_p \in \mathbb{R}^{P_1\cdot P_2\cdot C}$. H, W, and C, respectively, represent the height, width, and number of channels of the input image. P_1 and P_2, respectively, represent the height and width of the patch. The number of patches is $N = \frac{HW}{P_1P_2}$. Then, the x_p is projected into a d dimension token sequence using a linear projection layer. Meanwhile, a learnable CLS token $x_c \in \mathbb{R}^{1\times d}$ is added to the first position of the sequence to represent the global information of the image. Next, a 1D positional embedding vector $P \in \mathbb{R}^{(N+1)\times d}$ is embedded into the token sequence to encode positional information. After that, the token sequence is input into L stacked transformer encoders for self-attention calculation. Finally, the first output representation which can achieve global information integration is usually used for classification.

The transformer encoder is a key component of the transformer architecture, consisting of multiple blocks that incorporate multihead self-attention (MSA) and a feed-forward network (FFN). The FFN is composed of two layers of multilayer perceptrons (MLPs), with a RELU activation function following the first linear layer. To ensure stable training and improve performance, layer normalization is applied before the input of each block. Additionally, residual connections are employed in the output of each block to facilitate the flow of information through the network and mitigate the vanishing gradient problem.

The proposed SA encoder of SAG-CET utilizes the same components as a ViT, with the key difference being that while images have height, width, and channel dimensions, brainwave signals have only two dimensions—time and channel number. Therefore, by reshaping the brainwave signals and adding a padding dimension, they can be fully adapted for use with the encoder component of the ViT.

Definition of graph: We define $G = (V, E, A)$ as a graph, where V is the set of nodes. Each element $v_i \in V^C$ in the set represents a node of the graph, where C represents the number of channels and the value of v_i represents the feature of the node. E is a set of edges used to represent the connections between nodes. $A \in R^{C\times C}$ is a learnable adjacency matrix used to represent the connectivity relationship between nodes. $a_{ij} = 1$ represents that node i and node j are mutually connected. If a self-loop $a_{ii} = 1$ is added to the adjacency matrix, it can be represented in matrix form as $\hat{A} = A + I$. The identity matrix is represented as $I \in R^{C\times C}$. Define D as the degree matrix, where the degree matrix is a diagonal matrix with values $D_i = \sum_{j\in C} a_{ij}$ on the diagonal and zero in other regions. In the case of the adjacency matrix containing self-loops, $\hat{D} = D + I$. Define the Laplacian operator to describe the difference between a node and its adjacent nodes in a graph. Its matrix form L is defined as follows:

$$\begin{aligned} \Delta f &= DX - AX = (D - A)X \\ L &= D - A \end{aligned} \tag{1}$$

The Laplacian matrix is a matrix representation of a graph, which can be used to capture the relationships between nodes in the graph structure, as shown in Figure 3. In order to ensure the weighted aggregation of the first-order neighbor information of

nodes, the weight is inversely proportional to the degree of the node. The normalized Laplacian matrix is defined as

$$L_{sym} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{2}$$

Define the graph Fourier transformation and the graph inverse Fourier transformation as follows:

$$\begin{aligned} \hat{x} &= U^T x \\ x &= U\hat{x} \end{aligned} \tag{3}$$

where U is the eigenvector matrix of the Laplacian matrix.

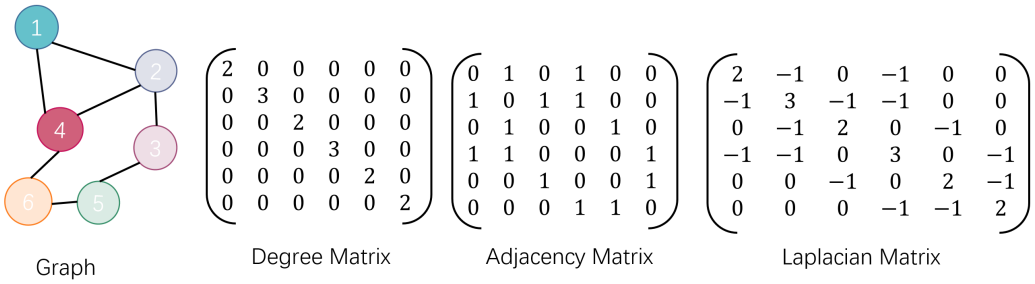

Figure 3. The Laplacian matrix represents the relationships within a graph.

3.2. Overview of Cascade Scale-Aware Adaptive Graph and Cross-Transformer Method

We propose a novel architecture named the cascade scale-aware adaptive graph and cross-transformer (SAG-CET) method. As illustrated in Figure 4, SAG-CET consists of two primary components: a scale-aware adaptive GCN (SAG) and a cross-EEG transformer (CET).

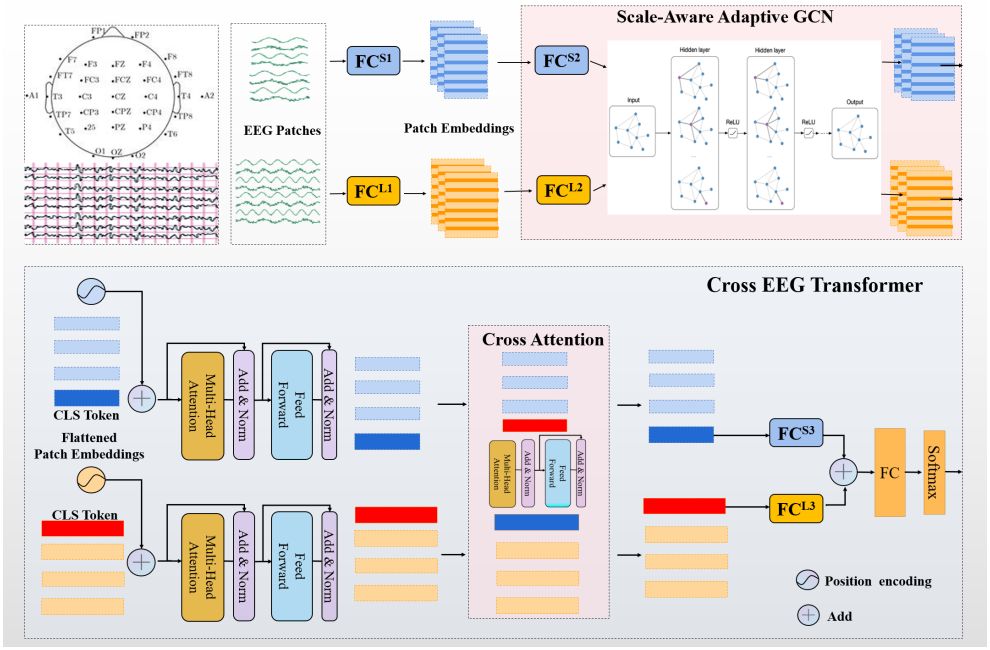

Figure 4. Overview of our SAG-CET. For each EEG sample, we first divide it into two kinds of multichannel patches and then use linear projection layers to obtain patch embeddings and further feed them into the proposed scale-aware adaptive GCN module, aiming to get channel-wise enhanced features by modeling the spatial relationship. Subsequently, we flatten patch embeddings and add a class token and position embedding for both short- and long-term patches. And then both patch features are fed into a self-attention encoder and a cross-attention module. Finally, both class token features are projected into the same dimension and averaged to form the final representation. The cross-entropy loss is used after Softmax for training.

First, we divide an EEG signal sample into two signal sequences according to two proportions and project them through the FC layer to form temporal patches. For an EEG sample with L length, the default proportions used to divide are L/32 and L/4. Then, we build a spatial electrode graph from the dual stream, where SAG is employed to learn the graph representation that captures the shared spatial correlation of multichannel connections between short and long temporal patches. After that, we use the self-attention encoder and the proposed novel cross-attention module to capture two global representations, i.e., class tokens, from two scale streams. And the class tokens are reprojected into the same dimension and averaged to a final representation for classification. Next, we will introduce the main components of our model in detail.

3.3. Scale-Aware Adaptive GCN

Considering the low spatial resolution of EEG signals and the interaction between multiple channels, it is imperative to account for the spatial correlation between these signals to enhance the performance of the model. Therefore, we propose the SAG to effectively learn and capture the spatial correlation present among various EEG channels. In the SAG, a learnable adjacency matrix $\hat{A}$ is used to construct the edge connections between nodes, and then a learnable weight matrix W is used to update the weights of the edges. The adjacency matrix $\hat{A}$ represents the edge connections, and the weight matrix W represents the edge weights, which together define the spatial relationships within the graph.

Spatial electrode graph construction: In the definition of the graph $G = (V, E, A)$, the nodes $v_i \in V^C$ of the spatial electrode graph represent EEG electrode channels, where v_i denotes the amplitude of the EEG signal at a single time step. The $\hat{A} \in R^{N \times N}$ is a learnable adjacency matrix, and it captures the connectivity between electrode channels in EEG signals. E is a set of edges that connect the EEG electrode channels, and the values of its elements are determined by $\hat{A}$.

Spatial electrode graph embedding: We learn the embedding of each node in the spatial electrode graph with the SAG, aiming to fully exploit the connections between the dual streams. First, the node sequences from each stream are projected using the functions $f(x)^l$ and $f(x)^s$ to align their dimensions. The dimension alignment functions $f(x)^l$ and $f(x)^s$ are fully connected layers that are used to transform the large-scale signal dimensions into small-scale signal dimensions. Subsequently, the resulting multiscale graph node sequence is fed into the SAG to compute the average weight. More specifically, the graph convolution process for the signals x and y is as follows:

$$x *_g y = U\big((U^T x) \odot (U^T y)\big) \tag{4}$$

where $\odot$ means the Hadamard product. Suppose $g(\cdot)$ is a filtering function (convolutional kernel); then, the signal x filtered by $g(L)$ can be expressed as

$$y = g(L)x = g(U\Lambda U^T)x = Ug(\Lambda)U^T x \tag{5}$$

where the convolutional kernel $g(\Lambda) = \mathrm{diag}([\hat{y_1}, \hat{y_2}, \cdots, \hat{y_n}])$ is a diagonal matrix. To simplify computations, we use K-order Chebyshev polynomials [46] to replace the polynomial expansion of $g(L)$. and we replace the convolution kernel with

$$g(\Lambda) = \sum_{k=0}^{K-1} \theta_k T_k(\Lambda) \tag{6}$$

Here, the coefficients of the Chebyshev polynomials are given by θ_k, and the eigenvector obtained by performing eigenvalue decomposition on the Laplacian matrix is $\Lambda = \mathrm{diag}([\lambda_1, \lambda_2, \cdots, \lambda_N])$. $T_k(\Lambda)$ can be calculated according to the following formula:

$$\begin{cases} T_0(x) = 1, T_1(x) = x \\ T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad k \geq 2 \end{cases} \tag{7}$$

According to (6), we can rewrite the graph convolution operation defined in (5) as

$$\begin{aligned} y &= Ug(\Lambda)U^T x \\ &= \sum_{k=0}^{K-1} U \begin{bmatrix} \theta_k T_k(\lambda_0) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_k T_k(\lambda_{N-1}) \end{bmatrix} U^T x \\ &= \sum_{k=0}^{K-1} \theta_k T_k(L) x \end{aligned} \tag{8}$$

3.4. Cross-EEG Transformer

Due to the physical sciences indicating that numerous physiological components function at multiple scales [30,31,47–49] and the ability of attention mechanisms to capture long-range temporal correlations, we propose the CET to integrate EEG signals of varying temporal scales.

In the initial stage, each stream undergoes spatial feature extraction via the SAG to initialize a learnable class token to represent global temporal features. Subsequently, each class token captures temporal features from other patches within the stream through the SA encoder. The CET incorporates multiple fusion and information exchange mechanisms within the cross-attention block, facilitating the class tokens to effectively capture multiscale features within the EEG signals. Finally, the class tokens obtained through the CET are utilized for emotion prediction.

The cross-attention block plays a pivotal role in information exchange between dual streams. Specifically, it facilitates the exchange of information between the class tokens and the temporal patches from the other stream, followed by the projection of the exchanged class tokens back to their respective streams. As a result, each stream is able to incorporate relevant information from the other stream, thereby significantly enhancing its ability to comprehend and handle multiscale inputs. Figure 5 illustrates the cross-attention block for the long stream. First, the class token of the long stream x_{cls}^{l} is subject to an alignment projection function denoted as $f^{l}(x)$, resulting in the transformed vector $x_{cls}^{\prime l}$ as defined in Equation (9). Subsequently, $x_{cls}^{\prime l}$ is utilized to generate the query by means of $W^{Q} \in R^{C\times(C/h)}$. Simultaneously, $x_{cls}^{\prime l}$ is concatenated with the temporal patches of the short stream x_{p}^{s} to yield $x^{\prime l}$, which is employed to derive the key and value via $W^{K} \in R^{C\times(C/h)}$ and $W^{V} \in R^{C\times(C/h)}$, respectively. The attention weights are calculated using the query, key, and value.

$$x_{cls}^{\prime l} = f^{l}\left(x_{cls}^{l}\right) \tag{9}$$

$$x^{\prime l} = \left[x_{cls}^{\prime l} \| x_{p}^{s}\right] \tag{10}$$

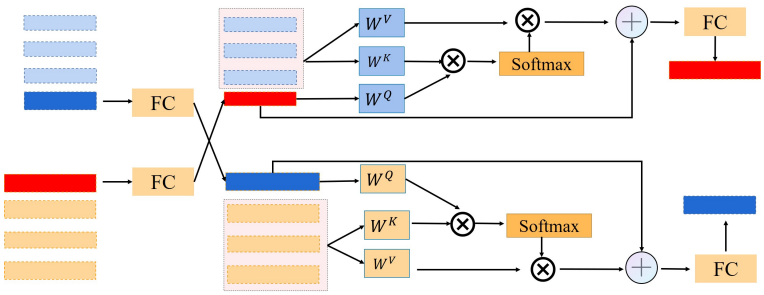

Figure 5. Illustration of the cross-attention module. Both short- and long-term class tokens are first projected to the same dimension of the other token, and then they integrate information from the other scale features by self-attention operation, and finally, each is reprojected into its original dimension.

The process of the cross-attention (CA) block can be expressed as

$$\begin{aligned} Q = x'^{l}_{cls} W^{Q}, K = x'^{l} W^{K}, V = x'^{l} W^{V} \\ A = \text{softmax}\left(QK^{T}/\sqrt{C/h}\right), \text{CA}\left(x'^{l}\right) = AV \end{aligned} \tag{11}$$

where C is the dimension of embedding and h is the number of heads. Cross-attention also uses multihead cross-attention (MCA) by referring to the SA mechanism, but unlike a vanilla unitary transformer, it does not use a feed-forward network after MCA. Specifically, the process of MCA can be represented as follows:

$$\begin{aligned} y^{l}_{cls} = f^{l}\left(x^{l}_{cls}\right) + \text{MCA}\left(\text{LN}\left(\left[f^{l}\left(x^{l}_{cls}\right) \| x^{s}_{p}\right]\right)\right) \\ z^{l} = \left[g^{l}\left(y^{l}_{cls}\right) \| x^{l}_{p}\right] \end{aligned} \tag{12}$$

where $g^{l}(x)$ is the inverse projection function, which is used to reproject the class token that has undergone information fusion back to its original dimension for concatenation with temporal patches. Finally, the pseudocode of the SAG-CET is shown in Algorithm 1.

Algorithm 1 Training procedure of SAG-CET

Input:raw signal data D
Output:Network $\theta^{(N)}$
Initialize:$\hat{A}^{(0)}, \mathbf{x}^{l(0)}_{cls}, \mathbf{x}^{s(0)}_{cls}, \theta^{(0)}$

$Patch_s, Patch_l = \text{SplitPatch}(D, s_{size}, l_{size})$
$Node_s, Node_l = \text{PatchToNode}(Patch_s, Patch_l)$
$NodeSet = \text{align\_cat}(Node_s, Node_l)$
for $i = 0 \ldots N-1$ **do**
  $Node'_s, Node'_l = \text{SAG}(NodeSet, \hat{A}^{(i)})$
  $Patch'_s, Patch_l = \text{NodeToPatch}(Node'_s, Node'_l)$
  $\mathbf{x}^{l(i)}_{cls}, \mathbf{x}^{s(i)}_{cls} = \text{CET}(Patch'_s, Patch'_l, \mathbf{x}^{l(i)}_{cls}, \mathbf{x}^{s(i)}_{cls})$
  $output = \text{classifier}(\mathbf{x}^{l(i)}_{cls} + \mathbf{x}^{s(i)}_{cls})$
  $AccSub = \text{acc}(label, output)$
  $loss = \text{nn.CrossEntropyLoss}(output, label)$
  $\hat{A}^{(i)}, \mathbf{x}^{l(i)}_{cls}, \mathbf{x}^{s(i)}_{cls} \leftarrow \text{loss.backward}()$
end for
return $\theta^{(N)}$

4. Experiments

4.1. Datasets and Experiment Setting

We conducted subject-dependent experiments on two widely used public datasets, namely, DEAP [50] and DREAMER [51], which have been extensively studied by researchers in the field of emotion recognition, as shown in Table 1. DEAP is a multimodal physiological signals dataset, which comprises recordings from 32 subjects. The dataset includes 32 channels of EEG signals and 8 channels of peripheral physiological signals (PPS), with an EEG signal sampling rate of 512 Hz. The remaining eight channels of PPS consist of signals for EMG, EOG, GSR, blood volume pulse (BVP), temperature, and respiration. The subjects were instructed to rate their valence, arousal, and dominance on a scale of 1 to 9 for the four emotional dimensions. DREAMER includes EEG signals from 14 channels and ECG signals from 2 channels of 23 subjects. The EEG signals were sampled at a rate of 128 Hz. The subjects were asked to rate four emotional dimensions, namely, valence, arousal, dominance, and liking, on a scale of 1 to 5.

Table 1. The details of DEAP and DREAMER.

Dataset	Subjects	Label	Experiments	Signals (Channel)
DEAP	32	Valence, arousal, dominance	40/subject	EEG (32), EMG (2), EOG (2), GSR (1), BVP (1), temperature (1), respiration (1)
DREAMER	23	Valence, arousal, dominance, liking	18/subject	EEG (14), ECG (2)

To ensure consistency and comparability between the two datasets, we downsampled all EEG signals to 128 Hz and extracted a 60-second segment from each signal. We filtered EEG signals to 4–45 Hz and performed blind source separation to remove EOG artifacts [50,51]. To enhance the characterization ability of the EEG signals, we performed baseline correction on the DEAP and DREAMER datasets by subtracting the average of each second for three and four seconds, respectively [52]. To increase the scale of our training set, we augmented the dataset by extracting nonoverlapping 1-s signal segments as samples [53]. Finally, following the common setting [6,54], we categorized the valence and arousal labels in the DEAP and DREAMER datasets as either high or low using thresholds of 5 and 3, respectively.

For each EEG sample with length L, we split it into long-term patches as L/4 and short-term patches L/32 by default. For our experiments, we set the initial learning rate to 0.001 and the batch size to 64. We utilized Adam as the optimizer and set the β_1 and β_2 parameters to 0.5 and 0.9, respectively. To ensure the reliability of our results, we employed a random 10-fold cross-validation approach and trained our model for 100 epochs for each fold. We evaluated the performance of our model using accuracy metrics. All experiments were conducted on Tesla A100 GPUs.

4.2. Comparisons to State of the Art

On DEAP and DREAMER, we compared our model with the SOTA methods, which use different backbones. The used models contain CRAM [20], ACRNN [42], MMResLSTM [55], DGCNN [56], IAG [26], V-IAG [29], SSTemotionNet [57], HetEmotionNet [43], EeT [54], and TNAS [6], as shown in Table 2. To ensure the comprehensiveness of the comparative experiments, these models adopt different feature selection strategies, including spatial–temporal domain features [20,55], spatial–spectral domain features [26,29,56], and the integration of spatial–spectral–temporal domain features [43,57]. Apart from this, we chose a wide range of models which include architectures featuring different combinations of CNN, RNN, LSTM, GRU, GCN, transformer, and others to comprehensively demonstrate the difference of our model compared to previous works. To ensure a fair comparison, we applied identical data preprocessing steps to both our model and the selected baseline models on both datasets and conducted comparative experiments under the same experiment conditions.

Table 2. Comparison of the input features and backbones in state-of-the-art methods.

Model	Feature	Modality	Backbone
CRAM [20]	Raw signal	EEG	CNN + LSTM + attention
SSTemotionNet [57]	Raw signal + DE	EEG	CNN + attention
ACRNN [42]	Raw signal	EEG	CNN + LSTM + attention
DGCNN [56]	PSD	EEG	GCN
IAG [26]	PSD	EEG + PPS	GCN + LSTM
V-IAG [29]	PSD	EEG + PPS	GCN + LSTM
HetEmotionNet [43]	Raw signal + DE	EEG + PPS	GCN + GRU
MMResLSTM [55]	Raw signal	EEG + PPS	LSTM + ResNet
EeT [54]	Raw signal	EEG	Transformer
TNAS [6]	Raw signal	EEG	Transformer
SAG-CET (ours)	Raw signal	EEG	GCN + transformer

4.3. Experiment on DEAP

Table 3 presents the average accuracy and standard deviations of the models on the DEAP dataset. Figure 6 shows the average accuracy per subject. Our model achieves an accuracy of 98.89% and 98.92% in valence and arousal, respectively, while the other baseline methods ranged between 84.46% and 98.68%. In these models, TNAS, which has a performance (98.66% for valence, 98.68% for arousal) similar to SAG-CET, combines the advantages of transformer and MOEA to build a backbone network. However, the performance of TNAS on DREAMER is far inferior to SAG-CET, indicating that SAG-CET has better generalization ability and model robustness than TNAS. Additionally, the proposed model demonstrates satisfactory stability, as evidenced by its lowest standard deviation. The results demonstrate that our proposed model achieves better performance than SOTA baselines on DEAP. Figure 7 provides a confusion matrix to illustrate the correctness of the model. The sample size of the DEAP dataset is $60 \times 40 \times 32$, and the sample size of the DREAMER dataset is $60 \times 23 \times 18$.

Table 3. The mean accuracies (ACC) and standard deviations (STD) on the DEAP dataset.

Model	Valence (%)	Arousal (%)
CRAM [20]	87.09 ± 7.49	84.46 ± 9.27
DGCNN [56]	90.44 ± 3.01	91.70 ± 3.46
MMResLSTM [55]	92.30 ± 1.55	92.87 ± 2.11
EeT [54]	93.34 ± 2.12	92.86 ± 2.35
ACRNN [42]	93.72 ± 3.21	93.38 ± 3.73
SST-EmotionNet [57]	95.54 ± 2.54	95.97 ± 2.86
HetEmotionNet [43]	97.66 ± 1.54	97.30 ± 1.65
TNAS [6]	98.66 ± 0.94	98.68 ± 0.98
SAG-CET (ours)	98.89 ± 0.84	98.92 ± 0.81

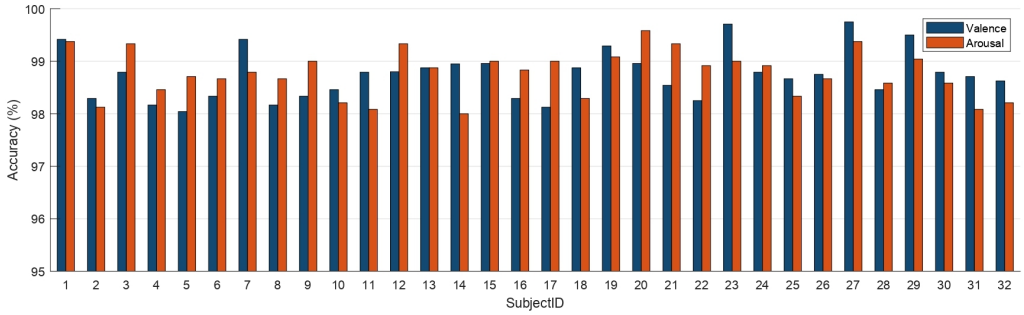

Figure 6. Subject-dependent classification accuracy on DEAP.

4.4. Experiment on DREAMER

Table 4 showcases the performance of the proposed model and other baseline models on the DREAMER dataset. Figure 8 shows the average accuracy per subject. In general, our SAG-CET achieves the best performance with an accuracy of (99.08% for valence and 99.21% for arousal) on the DREAMER dataset compared to all previous SOTA methods, with the lowest standard deviation.

Notably, although ACRNN seems to have achieved a slightly lower performance (97.39% for valence, 97.98% for arousal) on the DREAMER than SAG-CET, compared to performance on the DREAMER dataset, SAG-CET far outperforms ACRNN on the DEAP dataset, which indicates that SAG-CET has excellent generalization ability.

Table 4. The mean accuracies (ACC) and standard deviations (STD) on the DREAMER dataset.

Model	Valence (%)	Arousal (%)
DGCNN [56]	86.23 ± 3.25	84.54 ± 3.16
IAG [26]	90.75 ± 2.27	91.03 ± 1.65
CRAM [20]	92.27 ± 2.95	93.03 ± 1.87
V-IAG [29]	92.82 ± 2.16	93.09 ± 1.44
TNAS [6]	96.95 ± 3.35	96.41 ± 3.61
ACRNN [42]	97.39 ± 1.37	97.98 ± 1.92
SAG-CET (ours)	99.08 ± 0.51	99.21 ± 0.72

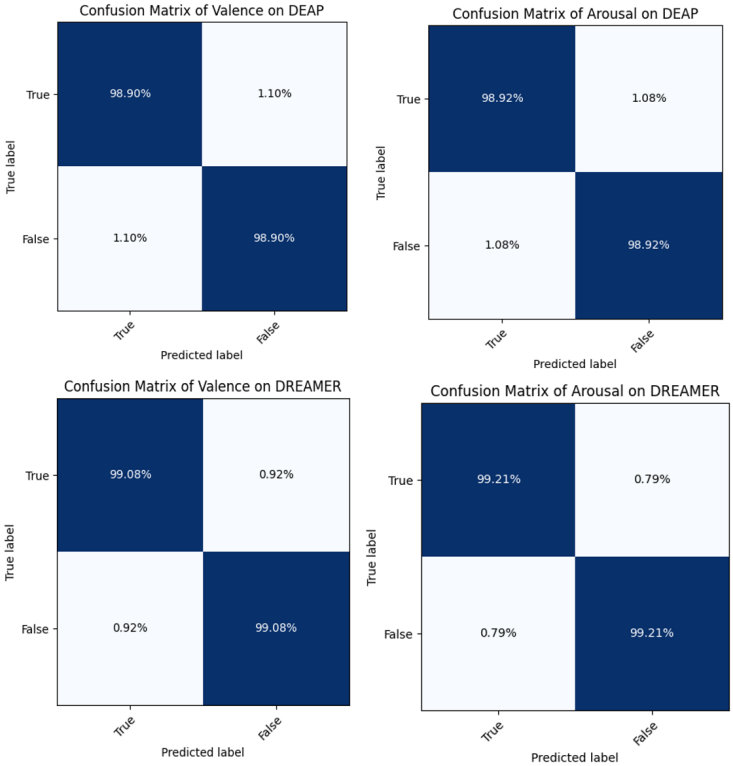

Figure 7. Accuracy fusion matrices of DEAP and DREAMER.

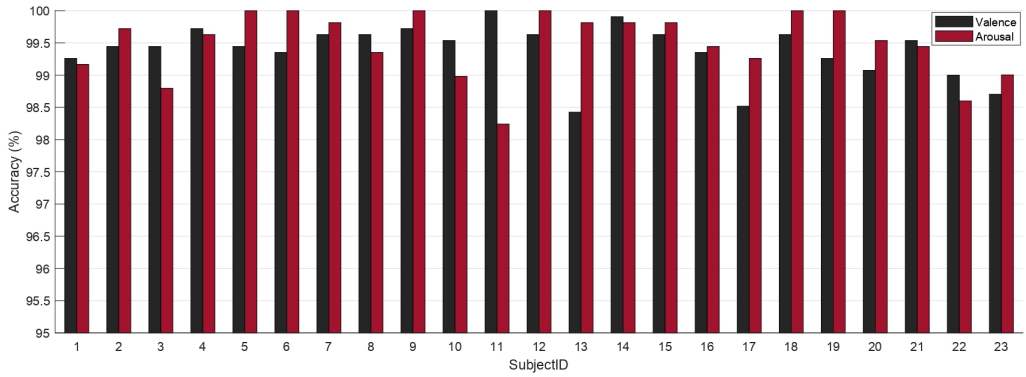

Figure 8. Subject-dependent classification accuracy on DREAMER.

4.5. Results Analysis

Figure 9 shows the average weight matrix of MHA of each stream in each subject before performing cross-attention calculation, which indicates the distribution of attention in the input embedding. The *i*-th row of each attention weight matrix corresponds to the attention distribution of class token of *i*-th subject over the entire embedding sequence. In the long stream, attention mainly focuses on the middle to the posterior part of temporal patches, while

in the short stream, attention focuses on the anterior part of temporal patches. This indicates that the features of the EEG signals in temporal patches at different scales are complementary.

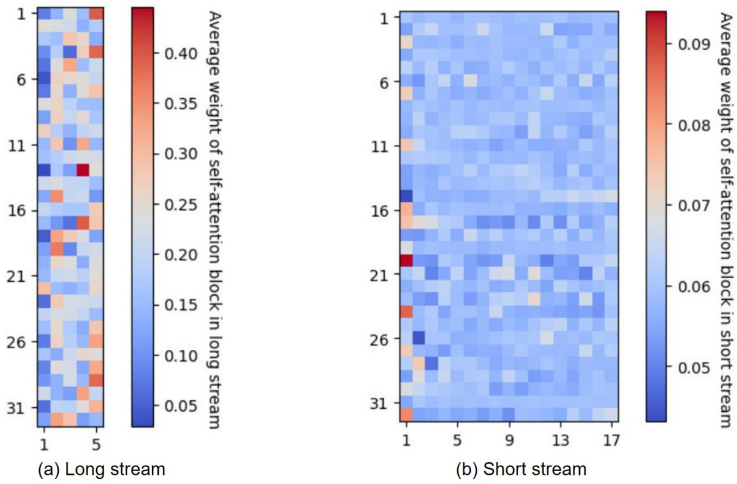

Figure 9. Example of average weight matrix of MHA head of each stream. The proportions of temporal patch size are 1/4 and 1/16 in (**a**) and (**b**), respectively. The *i*-th row of each attention weight matrix corresponds to the attention distribution of class token of *i*-th subject over the entire embedding sequence.

To more intuitively describe the ability of EEG signals to represent emotion at various scales, we explored the impact of different patch sizes on the performance of EEG emotion recognition using the single-stream input transformer. As shown in Figure 10, results indicate that the accuracy gradually improves with the increase in patch size. However, it is worth noting that the growth rate of accuracy is negatively correlated with the patch size. This suggests that although larger patch sizes can capture more temporal information, there is a limit to the improvement in accuracy through this approach. In order to find the optimal scale for using SAG-CET for EEG emotion recognition, we conducted a set of experiments to evaluate the optimal combination of patch proportion in a segment of an EEG signal sample. As shown in Figure 10, when the patch sizes for short-stream and long-stream inputs are, respectively, 1/32 and 1/4 of the EEG signal sample, CET achieves the best performance on both datasets. Furthermore, our proposed CET model overcomes the limitation shown in Figure 10 by fusing features from two different patch sizes, demonstrating the effectiveness of multiscale features in EEG emotion recognition.

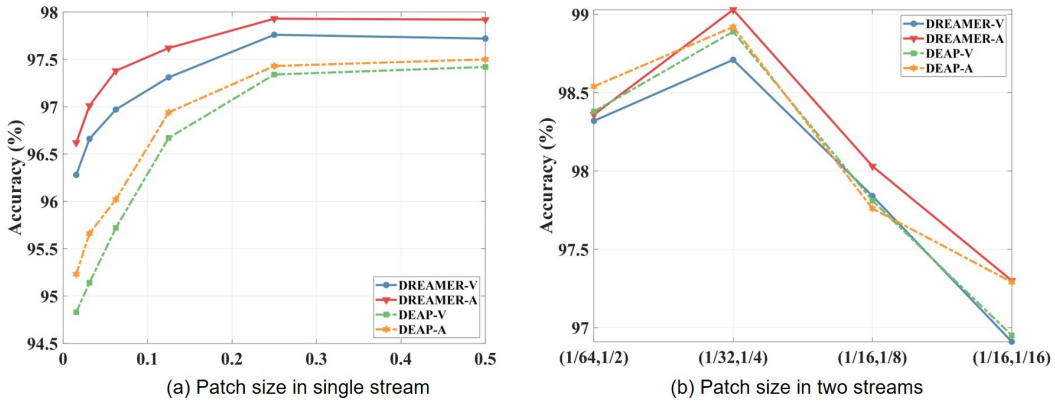

Figure 10. (**a**) The horizontal axis represents the proportion of patch size in a single sample, and the proportions selected for this experiment are 1/64, 1/32, 1/16, 1/8, 1/4, and 1/2. (**b**) The coordinates pair (s, l) on the horizontal axis represents the combinations of patch size proportion in a single sample for short and long streams.

4.6. Ablation Study

In order to verify the effectiveness of the proposed model, we conducted two types of ablation studies on two datasets: (1) Ablation studies to validate the effectiveness of the SAG in extracting spatial features. (2) Ablation studies to verify if the dual-stream structure and CET can effectively perform multiscale feature fusion. The single-stream input only uses an SA encoder, and the dual-stream input additionally uses a cross-attention block. The results are as shown in Figures 11 and 12; we have the following observations:

- By comparing the results of experiments conducted solely using long-stream, short-stream, or dual-stream input on two datasets, we can observe that dual-stream input performs better than long-stream or short-stream input, which fully explains the effective promotion with the application of the cross-attention block.
- By comparing the results of experiments conducted on the SAG block, removing the SAG component from the SAG-CET reduces the performance. The CET with SAG outperforms the CET without SAG, which indicates that the SAG is effective in constructing the spatial correlations of multiple channels. Notably, our analysis showed that the best performance was achieved by combining all the aforementioned factors across all the experiments.

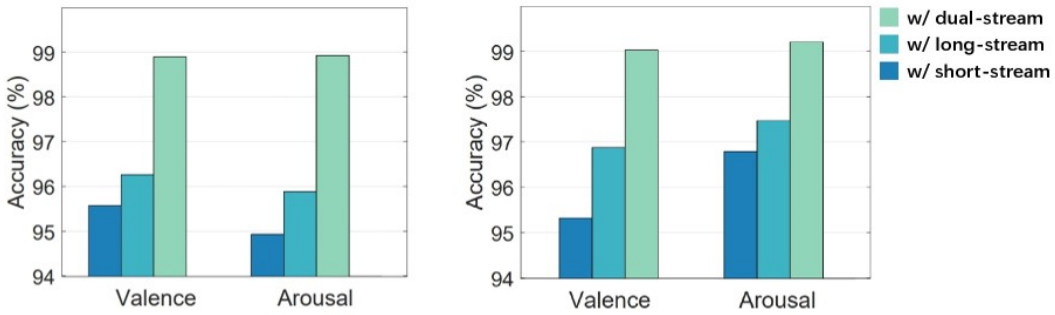

Figure 11. Ablation studies for CET structure on DEAP (**left**) and DREAMER (**right**).

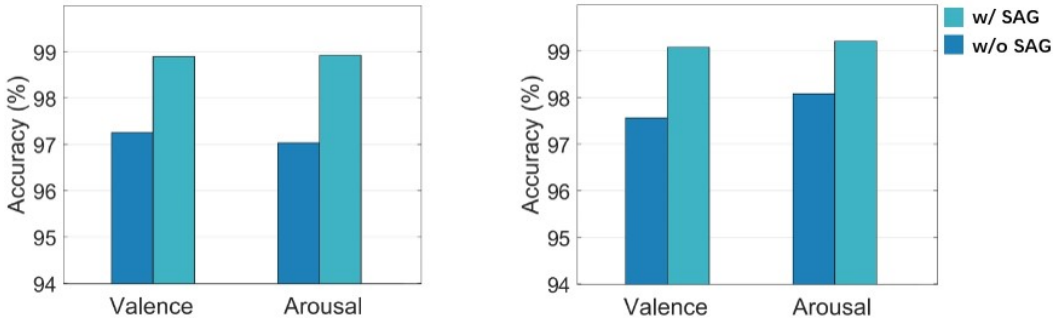

Figure 12. Ablation studies for SAG structure on DEAP (**left**) and DREAMER (**right**).

4.7. Limitations and Future Work

In this study, as a common setting, we adopt a conventional binary classification to classify the valence and arousal of emotions and only conduct subject-dependent experiments. Thus, the generalization of a subject-independent setting is unknown, and multiclass tasks on valence and arousal can be further explored. In addition, since EEG can represent extensive physiological states, another future work is to adapt our method to other states like mental fatigue as well as sleep monitoring, epilepsy monitoring, depression treatment, and other aspects.

5. Conclusions

In this paper, we propose a novel approach to EEG signal emotion recognition using end-to-end deep learning with an emphasis on multiscale and multichannel feature extraction. The proposed model, SAG-CET, utilizes a scale-aware adaptive GCN, dual-stream

input with varying scale EEG signal patches, and a cross-attention block that enables the fusion of features from signals of different scales. Extensive experiments comparing our approach with various backbone network models on the DEAP and DREAMER datasets demonstrated that our model achieves SOTA performance, outperforming existing methods by a significant margin. Additionally, we conducted experiments to demonstrate the complementary nature of signal patches at different temporal segmentation scales. Through extensive experimentation, we discovered the optimal combinations of multiple scale ratios, contributing to the development of a multiscale segmentation strategy in the temporal dimension for EEG signals. Finally, we performed ablation experiments by removing two key modules to demonstrate the effectiveness of the proposed modules.

Author Contributions: Conceptualization, Funding acquisition, methodology, writing—review and editing, X.P.; conceptualization, formal analysis, methodology, validation, writing—original draft, writing—review and editing, Y.G.; writing—review and editing, X.F., B.Z. and X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the National Natural Science Foundation of China (62176165), the Stable Support Projects for Shenzhen Higher Education Institutions (20220718110918001), the Natural Science Foundation of Top Talent of SZTU (GDRC202131), the Basic and Applied Basic Research Project of Guangdong Province (2022B1515130009), and the Special subject on Agriculture and Social Development, Key Research and Development Plan in Guangzhou (2023B03J0172).

Data Availability Statement: The data will be made available by the authors on request.

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

References

1. Zheng, L.; Ma, Y.; Li, M.; Xiao, Y.; Feng, W.; Wu, X. Time-frequency decomposition-based weighted ensemble learning for motor imagery EEG classification. In Proceedings of the 2021 IEEE International Conference on Real-time Computing and Robotics (RCAR), Qinghai, China, 15–19 July 2021; IEEE: Piscateville, NJ, USA, 2021; pp. 620–625.
2. Tang, X.; Zhang, J.; Qi, Y.; Liu, K.; Li, R.; Wang, H. A Spatial Filter Temporal Graph Convolutional Network for decoding motor imagery EEG signals. *Expert Syst. Appl.* **2024**, *238*, 121915. [CrossRef]
3. Wang, Y.; Wu, Q.; Wang, S.; Fang, X.; Ruan, Q. MI-EEG: Generalized model based on mutual information for EEG emotion recognition without adversarial training. *Expert Syst. Appl.* **2024**, *244*, 122777. [CrossRef]
4. Mukherjee, P.; Roy, A.H. EEG sensor driven assistive device for elbow and finger rehabilitation using deep learning. *Expert Syst. Appl.* **2023**, *244*, 122954. [CrossRef]
5. Yu, C.; Wang, M. Survey of emotion recognition methods using EEG information. *Cogn. Robot.* **2022**, *2*, 132–146. [CrossRef]
6. Li, X.; Zhang, Y.; Tiwari, P.; Song, D.; Hu, B.; Yang, M.; Zhao, Z.; Kumar, N.; Marttinen, P. EEG based emotion recognition: A tutorial and review. *ACM Comput. Surv.* **2022**, *55*, 1–57. [CrossRef]
7. Ezzameli, K.; Mahersia, H. Emotion recognition from unimodal to multimodal analysis: A review. *Inf. Fusion* **2023**, *99*, 101847. [CrossRef]
8. Cacioppo, J.T. Feelings and emotions: Roles for electrophysiological markers. *Biol. Psychol.* **2004**, *67*, 235–243. [CrossRef]
9. Liu, Y.; Sourina, O.; Nguyen, M.K. Real-time EEG-based emotion recognition and its applications. In *Transactions on Computational Science XII: Special Issue on Cyberworlds*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 256–277.
10. Jie, X.; Cao, R.; Li, L. Emotion recognition based on the sample entropy of EEG. *Bio-Med. Mater. Eng.* **2014**, *24*, 1185–1192. [CrossRef]
11. Liu, Y.; Sourina, O. Real-Time Fractal-Based Valence Level Recognition from EEG. In *Transactions on Computational Science XVIII*; Gavrilova, M.L., Tan, C.J.K., Kuijper, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 101–120.
12. Shi, L.C.; Jiao, Y.Y.; Lu, B.L. Differential entropy feature for EEG-based vigilance estimation. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; IEEE: Piscateville, NJ, USA, 2013; pp. 6627–6630.
13. Frantzidis, C.A.; Bratsas, C.; Papadelis, C.L.; Konstantinidis, E.I.; Pappas, C.; Bamidis, P.D. Toward emotion aware computing: An integrated approach using multichannel neurophysiological recordings and affective visual stimuli. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 589–597. [CrossRef] [PubMed]
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Saurous, R.A. Tacotron: Towards End-to-End Speech Synthesis. *arXiv* **2017**, arXiv:1703.10135.

16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; NIPS'17, pp. 6000–6010.
17. Lin, L.; Li, P.; Wang, Q.; Bai, B.; Cui, R.; Yu, Z.; Gao, D.; Zhang, Y. An EEG-based cross-subject interpretable CNN for game player expertise level classification. *Expert Syst. Appl.* **2024**, *237*, 121658. [CrossRef]
18. Choo, S.; Park, H.; Kim, S.; Park, D.; Jung, J.Y.; Lee, S.; Nam, C.S. Effectiveness of multi-task deep learning framework for EEG-based emotion and context recognition. *Expert Syst. Appl.* **2023**, *227*, 120348. [CrossRef]
19. Srinivasan, S.; Johnson, S.D. A Novel Approach to Schizophrenia Detection: Optimized Preprocessing and Deep Learning Analysis of Multichannel EEG Data. *Expert Syst. Appl.* **2023**, *246*, 122937. [CrossRef]
20. Zhang, D.; Yao, L.; Chen, K.; Monaghan, J. A convolutional recurrent attention model for subject-independent EEG signal analysis. *IEEE Signal Process. Lett.* **2019**, *26*, 715–719. [CrossRef]
21. Li, Y.; Zheng, W.; Wang, L.; Zong, Y.; Cui, Z. From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* **2019**, *13*, 568–578. [CrossRef]
22. Du, X.; Ma, C.; Zhang, G.; Li, J.; Lai, Y.K.; Zhao, G.; Deng, X.; Liu, Y.J.; Wang, H. An efficient LSTM network for emotion recognition from multichannel EEG signals. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1528–1540. [CrossRef]
23. Bagchi, S.; Bathula, D.R. EEG-ConvTransformer for single-trial EEG-based visual stimulus classification. *Pattern Recognit.* **2022**, *129*, 108757. [CrossRef]
24. Shawky, E.; El-Khoribi, R.; Shoman, M.A.I.; Wahby, M.A. EEG-Based Emotion Recognition using 3D Convolutional Neural Networks. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 329.
25. Zhang, Z.; Meng, Q.; Jin, L.; Wang, H.; Hou, H. A novel EEG-based graph convolution network for depression detection: Incorporating secondary subject partitioning and attention mechanism. *Expert Syst. Appl.* **2024**, *239*, 122356. [CrossRef]
26. Song, T.; Liu, S.; Zheng, W.; Zong, Y.; Cui, Z. Instance-adaptive graph for EEG emotion recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 2701–2708.
27. Zhang, G.; Yu, M.; Liu, Y.J.; Zhao, G.; Zhang, D.; Zheng, W. SparseDGCNN: Recognizing Emotion From Multichannel EEG Signals. *IEEE Trans. Affect. Comput.* **2023**, *14*, 537–548. [CrossRef]
28. Jiang, W.B.; Yan, X.; Zheng, W.L.; Lu, B.L. Elastic Graph Transformer Networks for EEG-Based Emotion Recognition. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscateville, NJ, USA, 2023; pp. 1–5.
29. Song, T.; Liu, S.; Zheng, W.; Zong, Y.; Cui, Z.; Li, Y.; Zhou, X. Variational Instance-Adaptive Graph for EEG Emotion Recognition. *IEEE Trans. Affect. Comput.* **2023**, *14*, 343–356. [CrossRef]
30. Costa, M.; Goldberger, A.L.; Peng, C.K. Multiscale entropy analysis of biological signals. *Phys. Rev. E* **2005**, *71*, 021906. [CrossRef] [PubMed]
31. Song, Z.; Deng, B.; Wei, X.; Cai, L.; Yu, H.; Wang, J.; Wang, R.; Chen, Y. Scale-specific effects: A report on multiscale analysis of acupunctured EEG in entropy and power. *Phys. A Stat. Mech. Its Appl.* **2018**, *492*, S0378437117311901. [CrossRef]
32. Wang, Z. Emotion Recognition Based on Multi-scale Convolutional Neural Network. In *Proceedings of the Data Mining and Big Data*; Tan, Y., Shi, Y., Eds.; Springer Nature: Singapore, 2022; pp. 152–164.
33. Hu, J.; Wang, C.; Jia, Q.; Bu, Q.; Sutcliffe, R.; Feng, J. ScalingNet: Extracting features from raw EEG data for emotion recognitionPy. *Neurocomputing* **2021**, *463*, 177–184.
34. Wang, Z.M.; Zhang, J.W.; He, Y.; Zhang, J. EEG emotion recognition using multichannel weighted multiscale permutation entropy. *Appl. Intell.* **2022**, *52*, 12064–12076. [CrossRef]
35. Russell, J.A. Core affect and the psychological construction of emotion. *Psychol. Rev.* **2003**, *110*, 145. [CrossRef]
36. Li, M.; Lu, B.L. Emotion classification based on gamma-band EEG. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 3–6 September 2009; IEEE: Piscateville, NJ, USA, 2009; pp. 1223–1226.
37. Bazgir, O.; Mohammadi, Z.; Habibi, S.A.H. Emotion recognition with machine learning using EEG signals. In Proceedings of the 2018 25th national and 3rd international iranian conference on biomedical engineering (ICBME), Qom, Iran, 29–30 November 2018; IEEE: Piscateville, NJ, USA, 2018; pp. 1–5.
38. Khare, S.K.; Bajaj, V. An evolutionary optimized variational mode decomposition for emotion recognition. *IEEE Sens. J.* **2020**, *21*, 2035–2042. [CrossRef]
39. Zhong, S.h.; Fares, A.; Jiang, J. An attentional-LSTM for improved classification of brain activities evoked by images. In Proceedings of the 27th ACM international conference on multimedia, Nice, France, 21–25 October 2019; pp. 1295–1303.
40. 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 2016 IEEE international conference on bioinformatics and biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; IEEE: Piscateville, NJ, USA, 2016; pp. 352–359.
41. Liao, J.; Zhong, Q.; Zhu, Y.; Cai, D. Multimodal physiological signal emotion recognition based on convolutional recurrent neural network. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Chennai, India, 16–17 September 2020; IOP Publishing: Bristol, UK, 2020; Volume 782, p. 032005.
42. 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**, *14*, 382–393. [CrossRef]

43. Jia, Z.; Lin, Y.; Wang, J.; Feng, Z.; Xie, X.; Chen, C. HetEmotionNet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In Proceedings of the 29th ACM International Conference on Multimedia, online, 15 July 2021; pp. 1047–1056.
44. Jomaa, M.E.S.H.; Van Bogaert, P.; Jrad, N.; Kadish, N.E.; Japaridze, N.; Siniatchkin, M.; Colominas, M.A.; Humeau-Heurtier, A. Multivariate improved weighted multiscale permutation entropy and its application on EEG data. *Biomed. Signal Process. Control* **2019**, *52*, 420–428. [CrossRef]
45. Su, Y.; Zhang, Z.; Li, X.; Zhang, B.; Ma, H. The multiscale 3D convolutional network for emotion recognition based on electroencephalogram. *Front. Neurosci.* **2022**, *16*, 872311. [CrossRef] [PubMed]
46. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3844–3852.
47. Wright, J.; Liley, D. Dynamics of the brain at global and microscopic scales: Neural networks and the EEG. *Behav. Brain Sci.* **1996**, *19*, 285–295. [CrossRef]
48. Nunez, P.L. Toward a quantitative description of large-scale neocortical dynamic function and EEG. *Behav. Brain Sci.* **2000**, *23*, 371–398. [CrossRef] [PubMed]
49. Hramov, A.E.; Koronovskii, A.A.; Makarov, V.A.; Pavlov, A.N.; Sitnikova, E. *Wavelets in Neuroscience*; Springer: Berlin/Heidelberg, Germany, 2015.
50. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31. [CrossRef]
51. 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] [PubMed]
52. Yang, Y.; Wu, Q.; Qiu, M.; Wang, Y.; Chen, X. Emotion recognition from multi-channel 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: Piscateville, NJ, USA, 2018; pp. 1–7.
53. Wang, X.W.; Nie, D.; Lu, B.L. Emotional state classification from EEG data using machine learning approach. *Neurocomputing* **2014**, *129*, 94–106. [CrossRef]
54. Liu, J.; Zhang, L.; Wu, H.; Zhao, H. Transformers for EEG emotion recognition. *arXiv* **2021**, arXiv:2110.06553.
55. 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.
56. 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]
57. Jia, Z.; Lin, Y.; Cai, X.; Chen, H.; Gou, H.; Wang, J. Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2909–2917.