IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 21 March 2023, accepted 5 April 2023, date of publication 10 April 2023, date of current version 14 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266117

RESEARCH ARTICLE

Bi-Branch Vision Transformer Network for EEG Emotion Recognition

WEI LU[1,2], TIEN-PING TAN[1], AND HUA MA[2]

[1]School of Computer Sciences, Universiti Sains Malaysia (USM), Penang 11800, Malaysia
[2]Henan High-Speed Railway Operation and Maintenance Engineering Research Center, Zhengzhou Railway Vocational and Technical College, Zhengzhou 451460, China

Corresponding author: Tien-Ping Tan (tienping@usm.my)

This work was supported by the Henan Provincial Science and Technology Research Project under Grant 232102240091.

ABSTRACT Electroencephalogram (EEG) signals have emerged as an important tool for emotion research due to their objective reflection of real emotional states. Deep learning-based EEG emotion classification algorithms have made encouraging progress, but existing models struggle with capturing long-range dependence and integrating temporal, frequency, and spatial domain features that limit their classification ability. To address these challenges, this study proposes a Bi-branch Vision Transformer- based EEG emotion recognition model, Bi-ViTNet, that integrates spatial-temporal and spatial-frequency feature representations. Specifically, Bi-ViTNet is composed of spatial-frequency feature extraction branch and spatial-temporal feature extraction branch that fuse spatial-frequency-temporal features in a unified framework. Each branch is composed of Linear Embedding and Transformer Encoder, which is used to extract spatial-frequency features and spatial-temporal features. Finally, fusion and classification are performed by the Fusion and Classification layer. Experiments on SEED and SEED-IV datasets demonstrate that Bi-ViTNet outperforms state-of-the-art baselines.

INDEX TERMS Affective computing, EEG-based emotion recognition, transformer.

I. INTRODUCTION

Emotion is a psychological and physiological response formed by sensing external and internal stimuli that influences human behavior and plays a significant role in daily life [1], [2], [3], [4], [5], [6], [7]. As one of the most important research topics in affective computing, emotion recognition has garnered increasing interest in recent years due to its wide range of potential applications in human-computer interaction [8], disease detection [9], [10], [11], fatigue driving [12], [13], [14], [15], mental workload estimation [16], [17], [18], [19], and cognitive neuroscience. In general, emotion recognition methods can be divided into two types depending on whether physiological or non-physiological signals is involved [20]. Non-physiological signals, including speech, posture, and facial expression [21], are external manifestations of human emotions. Physiological signals, corresponding to the physiological reactions caused by emotions, such as eye electricity, ECG, EMG, and EEG, are human recessive emotional expressions [22]. Non-physiological signals, such as facial expressions and speech, are limited in their ability to reliably reflect an individual's true emotional state, because humans may conceal their emotions through masking their facial expression and voice. Physiological signals, on the other hand, are difficult to disguise and can objectively reflect human emotions. Consequently, physiological signals are more suitable for emotion recognition. Among physiological signals, EEG signals are characterized by high time resolution and rich information content, enabling the detection of subtle changes in emotions. These features make emotion recognition tasks based on EEG signals more objective and accurate than those based on other types of physiological signals [23], [24], [25], [26], [27]. Therefore, emotion recognition methods based on EEG signals are more favored by researchers.

In order to better complete emotion recognition based on EEG signals, it is necessary to extract multi-dimensional features of EEG signals. In general, EEG signal features can be divided into: temporal domain features of EEG

The associate editor coordinating the review of this manuscript and approving it for publication was Hui Ma.

signals, frequency domain features of EEG signals, and spatial domain features of EEG signals [28], [29]. *Temporal domain features of EEG signals*: EEG signals collected with scalp electrodes and EEG signals amplified with acquisition equipment are all expressed in temporal domain. In temporal domain analysis, the temporal changes of neurophysiological signals are used as features to describe EEG signals with precise time markers. Such features include the information extracted from the collected EEG signal and related to the peak value or duration, reflecting the change of the signal with time [30]. *Frequency domain features of EEG signals*: The collected EEG signals can be represented in frequency domain by Fourier transform or wavelet transform. In the frequency domain, the features of neural signals are the subband power of EEG signals and the Power Spectral Density (PSD) that reflects the power changes of specific EEG signal frequency bands [31]. *Spatial domain features of EEG signals*: The purpose of extracting spatial domain features of EEG signals is to identify brain regions that generate specific neural activities by drawing topographical map of brain [32], [33].

There are two main types of emotion recognition tasks based on EEG features [34]: conventional machine learning methods and deep learning methods. Conventional emotion recognition methods based on machine learning usually extract features from EEG signals, and then input these features into classification algorithm [35] such as Support Vector Machine (SVM) [36], K-Nearest Neighbor (KNN) [37], and Bayesian Network (BN) [38]. However, these methods require expert knowledge in both feature design and feature selection [39]. Therefore, it is challenging to extract relevant features from complex EEG signals and the results produced are lower compared to deep learning methods [40]. The effectiveness of deep learning in solving pattern recognition problems in natural language processing, computer vision [21], speech recognition, and other fields [41], [42], have inspired researchers to apply deep learning methods in emotion recognition tasks using for example Convolutional Neural Networks(CNN) [43] and Long Short-Term Memory (LSTM) [44]. Although these methods have led to improvements in emotion recognition results compared to conventional machine learning methods, these methods still face some challenges. *Challenge 1*: At present, most of the models do not integrate the features of the three different domains of EEG signal, namely spatial-frequency- temporal, which limits the classification ability of the models to some extent. *Challenge 2*: At present, most models do not have strong ability to capture long-range dependency, and it is difficult to capture the global information of EEG signals, thus extracting more powerful features, which affects the performance of model classification.

The aim of our research is to address the aforementioned challenges and improve the classification performance of the model. Therefore, in order to address the aforementioned challenges, we propose an EEG emotion recognition model, Bi-branch Vision Transformer Network (Bi-ViTNet), which is based on the dual branch Vision Transformer and takes spatial-frequency features representation and spatial-temporal features representation as input. Bi-ViTNet consists of spatial-frequency feature extraction branch and spatial-temporal feature extraction branch. Each branch is composed of a Linear Embedding and a Transformer Encoder, which is used to extract spatial-frequency features and spatial-temporal features. The extracted spatial-frequency features and spatial-temporal features are fused and classified by Fusion and Classification layer. Bi-ViTNet not only integrates the frequency-spatial-temporal information of EEG signals in a unified network framework but also the Transformer Encoder in each branch of Bi-ViTNet can better capture the long-range dependencies of EEG signals. Experiments using SEED and SEED-IV datasets show that Bi-ViTNet outperforms all the state-of-the-art models in terms of accuracy and standard deviation. Finally, we conduct ablation studies to determine the validity of each branching model.

II. RELATED WORK

We have reviewed related work in terms of EEG signal-based emotion recognition and the Transformer model in this section.

A. EEG-BASED EMOTION RECOGNITION

In recent years, time series data mining has gradually become a research hotspot [45], [46]. Time series technology has been applied in many fields, such as transportation [47] and medical treatment [48], [49], [50]. EEG is a typical time series data. EEG signals have been widely used in emotion recognition because they could reflect the real emotions of subjects accurately and objectively. In earlier studies, researchers used conventional machine learning models, such as SVM to model emotion using EEG signals. For example, Nie et al. extracted EEG features from the EEG signal, and employed a linear dynamic system technique to smooth these features, then modelled these features using SVM [51]. Anh et al. developed a real-time emotion recognition system based on EEG signals that is capable of detecting various emotional states, including happiness, relaxation, and neutral states. The system employs an SVM classifier and has demonstrated an average accuracy of 70.5% [52].

Inspired by the success of deep learning in computer vision, natural language processing, and biomedical signal processing, several researchers have attempted to employ deep learning methods for EEG-based emotion recognition. Zheng et al. proposed a deep belief network (DBN) to classify three categories of emotions using EEG features, and demonstrated through experiments that deep learning methods outperformed traditional machine learning methods [53]. Alhagry et al. put forward a kind of end-to-end deep learning neural network to recognize emotions from raw EEG signals. The network utilizes an LSTM-RNN to learn features from the EEG signals and the dense layer for classification [44].

Even though all the deep learning methods gave encouraging results, it is still difficult to combine more essential information from diverse domains. Therefore, some researchers proposed new methods. Al-Nafjan et al. proposed a methodology for EEG emotion recognition. Power spectral density and deep neural networks were considered in the proposed approach [54]. Yin et al. extracted Differential Entropy(DE) features to construct feature cubes and took them as input to a novel deep learning model as fusing graph convolutional neural network (GCNN) and LSTM to achieve EEG-based emotion classification [55]. Liu et al. developed a dynamic differential entropy (DDE) technique to extract EEG signal characteristics. The collected DDE features were input to a convolutional neural network [56]. Rahman et al. proposed to transform EEG signals into a topographic map of brain covering frequency and it was used as features to a convolutional neural networks for emotion recognition [57]. Topic et al. came up with an idea to construct topographic feature map (TOPO-FM) and holographic feature map (HOLO-FM) using EEG signal features and after that used deep learning as a feature extraction method on feature maps to identify different types of emotions [58].

These works obtained encouraging results using only one or two types of features (temporal or frequency features). There are few studies in EEG-based emotion recognition that combine three features: temporal domain features, frequency domain features, and spatial domain features. Jia et al. proposed an attention 3D dense network with fusing short-range EEG features in the time domain, the spatial domain, and the frequency domain [39]. Xiao et al. used 4D spatial-spectral-temporal representations as input and proposed a method called the 4D local attention-based neural network for EEG emotion classification and recognition [59]. However, these models do not learn long-range dependencies well and have difficulty in capturing the global information of EEG signals.

B. TRANSFORMER

Transformer was first proposed in natural language processing [60], and since then it has been applied in other domains [61]. Vision Transformer (ViT) based on multi-head self-attention to patches of images has achieved outstanding results in the field of computer vision. A ViT associates a query and a set of key-value pairs with an output based on the attention mechanism described as Formula (1):

$$Attention(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{D_k}})V \tag{1}$$

where Q is the query, K means the key, and V indicates the value, respectively. D_k represents the dimension of the query and the key. Through the training of large-scale data sets, Vision Transformer has achieved state-of-the-art ImageNet image classification result. In addition, it has been applied in computer vision problems, including object detection [62], image classification [63], segmentation [64], etc. In general, Transformer has the following advantages, *Advantage 1*: Transformer has strong ability to learn long-range dependencies, and its multi-head self-attention and parallel input processing improve the modeling of long-range dependencies. Transformer could use attention mechanisms to capture global information and extract more powerful features. *Advantage 2*: The position embedding of Transformer preserved key position information of words and image blocks, while class tags can aggregate representative information [65]. *Advantage 3*: Transformer is suitable for multi-modal data input. The data can be used as the input of Transformer model when converted into vectors. These advantages of Transformer model motivate us to investigate it for emotion recognition tasks based on EEG signals.

III. PRELIMINARIES

In this paper, we define $E^S = (E_1^S, E_2^S, \ldots, E_B^S) \in \mathbb{R}^{N_e \times B}$ as the frequency features containing B frequency bands extracted from EEG signals, where N_e is the number of electrodes. We construct the spatial-frequency features $A^S = (A_1^S, A_2^S, \ldots, A_B^S) \in \mathbb{R}^{H \times W \times B}$, where H and W represent the height and width of the frequency map, respectively.

We define $E^T = (E_1^T, E_2^T, \ldots, E_T^T) \in \mathbb{R}^{N_e \times T}$ as an EEG signal sample of T period, where N_e is the number of electrodes. We construct the spatial-temporal features $A^T = (A_1^T, A_2^T, \ldots, A_T^T) \in \mathbb{R}^{H \times W \times T}$, where H and W represent the height and width of the temporal map, respectively.

The objective of the study is to establish a mapping between spatial-frequency/temporal representations and emotional states. Given spatial-frequency representation A^S and spatial-temporal representation A^T, the emotion recognition task can be characterized as $Y_{out} = F(A^S, A^T)$, where Y_{out} represents the emotion state and and F is our proposed model.

IV. METHODOLOGY

We propose an EEG emotion recognition model, Bi-branch Vision Transformer Network (Bi-ViTNet), based on the dual branch Vision Transformer, with the spatial-frequency features representation and spatial-temporal features representation as the input. Figure 1 shows the overall architecture of the proposed Bi-ViTNet model. Bi-ViTNet consists of spatial-temporal encoder and spatial-frequency encoder, and inputs to the encoders are the spatial-frequency features and the spatial-temporal features respectively. Both spatial-temporal encoder and spatial-frequency encoder consist of a Linear Embedding Layer and a Transformer Encoder. The input features go into the Linear Embedding. We summarize three core ideas of the Bi-ViTNet as follows:

1) Spatial-frequency data construction and spatial-temporal data construction methods are proposed;
2) Based on the construction of spatial-frequency data and spatial-temporal data, the spatial-frequency-temporal information of EEG is fused in a unified network framework;

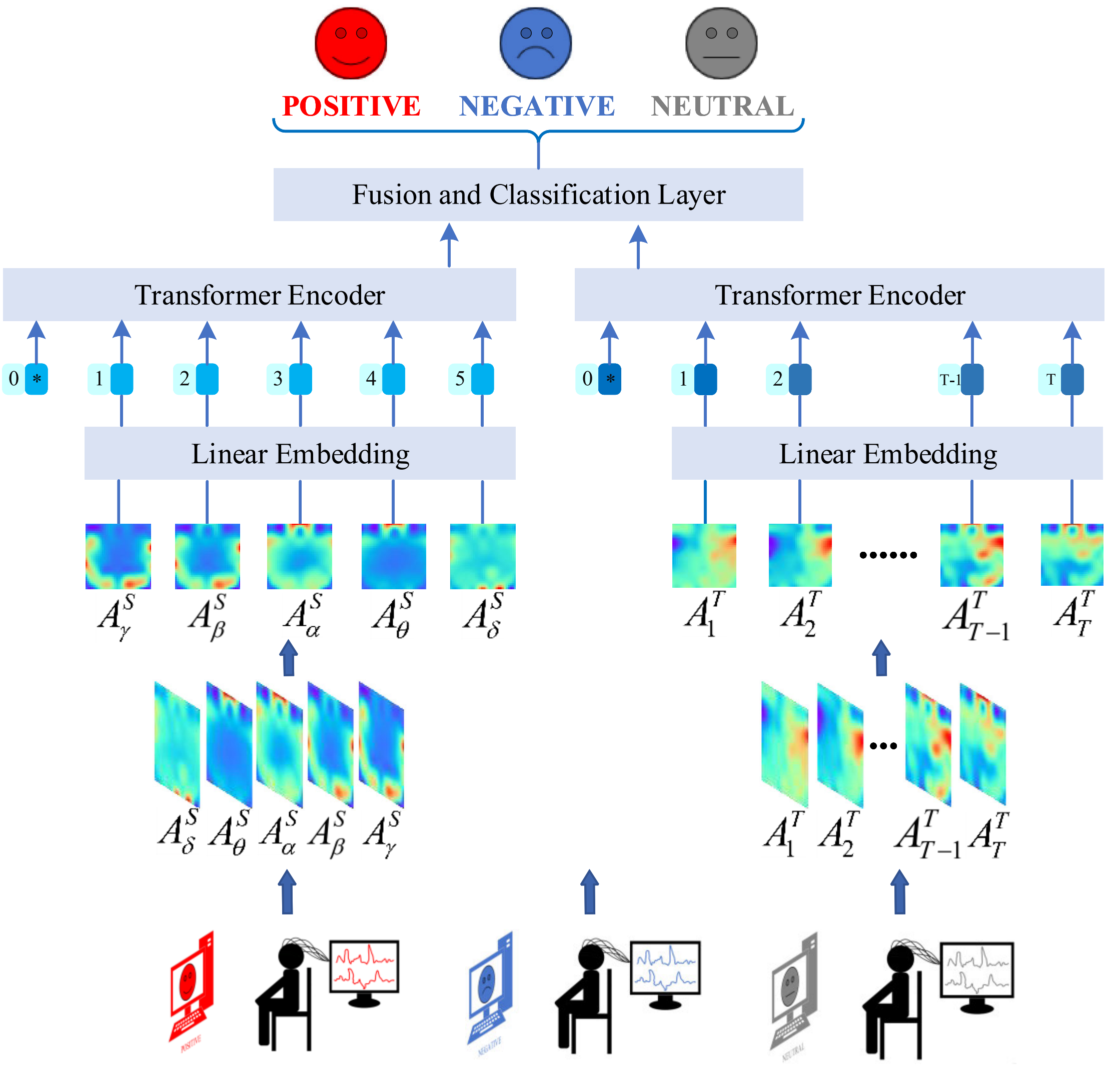

POSITIVE
NEGATIVE
NEUTRAL
Fusion and Classification Layer
Transformer Encoder
Transformer Encoder
0 *
1
2
3
4
5
0 *
1
2
T-1
T
Linear Embedding
Linear Embedding
A_γ^S A_β^S A_α^S A_θ^S A_δ^S
A_1^T A_2^T ...... A_{T-1}^T A_T^T
A_δ^S A_θ^S A_α^S A_β^S A_γ^S
A_1^T A_2^T ... A_{T-1}^T A_T^T
POSITIVE
NEGATIVE
NEUTRAL

FIGURE 1. **Whole process of EEG emotion recognition. The EEG signals of the experimental subjects are constructed as spatial-temporal and spatial- frequency representations and used as the input of the Bi-ViTNet model. The model consists of two branches: one branch extracts spatial-temporal fe atures and the other extracts spatial-frequency features. Both branches with the same structure are composed of Linear Embedding layer and Transfo rmer Encoder layer. Finally, the features extracted from the spatial-temporal branch and from the spatial-frequency branch are fused and classified.**

3) The Transformer model is used to capture the global information of EEG signals in spatial-temporal and spatial-frequency domains.

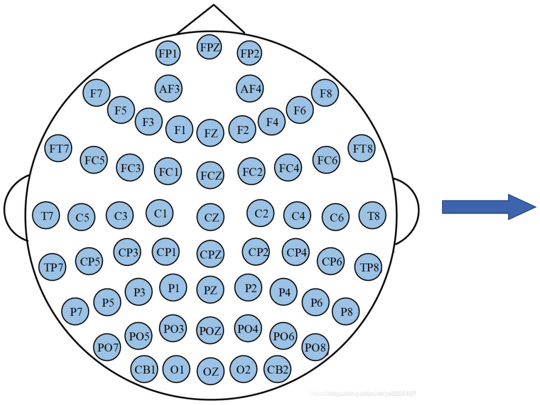
A. SPATIAL-FREQUENCY AND SPATIAL-TEMPORAL FEATURE REPRESENTATIONS

To better characterize EEG data, the original EEG signal is converted into spatial-temporal and spatial-frequency representations, which are used to describe the spatial distribution of temporal and frequency information of EEG signals. The 10-20 electrode placement system is an arrangement of electrodes on the surface, containing the spatial information of brain potential distribution. Then, the mapping electrode position matrix is used to create a spatial frame for each sample and describe the spatial information in the constructed spatial-temporal and spatial-frequency representations of EEG signals. The spatial-temporal and spatial-frequency representations of EEG signals are used as the input of the Bi-ViTNet, as shown in Figure 2.

Figure 3 shows the process of converting the original EEG signals into spatial-temporal and spatial-frequency representations. The original EEG signals are divided into non-overlapping periods lasting for $\tau = 1$ seconds, and each segment is assigned the same label as the original EEG signals.

1) SPATIAL-TEMPORAL FEATURE REPRESENTATION

To construct the spatial-temporal feature representation, we extract temporal-domain features of different time stamps from EEG fragments with a length of $\tau = 1$ seconds. We define $E^T = (E_1^T, E_2^T, \ldots, E_T^T) \in \mathbb{R}^{N_e \times T}$ as the EEG signal sample containing time stamp T, where the



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

FIGURE 2. Mapping of the Scalp Electrode Positions to a Matrix. The objective of this work is to maintain the positional links between several electrodes.

time stamp is $T \in \{1, 2, \ldots, 25\}$, the electrode is $N_e \in \{FP1, FPZ, \ldots, CB2\}$, and $X_t^T = (x_t^1, x_t^2, \ldots, x_t^N) \in \mathbb{R}^N (t \in \{1, 2, \ldots, T\})$ represent the EEG signals of all N electrodes collected on the time stamp T. Then, the selected data are mapped to a temporal-domain brain electrode position matrix $A_t^T \in \mathbb{R}^{H\times W} (t \in \{1, 2, \ldots, T\})$ according to the electrode location on the brain. Finally, the temporal-domain brain electrode position matrices from different time stamps are superimposed to form the spatial-temporal features representation of EEG, that is, the construction of $A^T = (A_1^T, A_2^T, \ldots, A_T^T) \in \mathbb{R}^{H\times W\times T}$ is completed.

2) SPATIAL-FREQUENCY FEATURE REPRESENTATION

To construct the spatial-frequency feature representation, the temporal-frequency feature extraction method is used to extract the Power Spectral Density features of five frequency bands $\{\delta, \theta, \alpha, \beta, \gamma\}$ of all EEG channels from the EEG signal samples in the EEG segments with a length of $\tau = 1$ seconds. We define $E^S = (E_1^S, E_2^S, \ldots, E_B^S) \in \mathbb{R}^{N_e\times B}$ as a frequency feature containing a frequency band extracted from the Power Spectral Density feature, in which the frequency band is $B \in \{\delta, \theta, \alpha, \beta, \gamma\}$, the electrode is $N_e \in \{FP1, FPZ, \ldots, CB2\}$, and $X_b^B = (x_b^1, x_b^2, \ldots, x_b^N) \in \mathbb{R}^N (b \in \{1, 2, \ldots, B\})$ represent the collection of EEG signals from all N_e electrodes on the frequency band B. Then, the selected data are mapped to a frequency domain brain electrode position matrix $A_b^S \in \mathbb{R}^{H\times W} (b \in \{1, 2, \ldots, B\})$ according to the electrode location on the brain. Finally, the frequency-domain brain electrode position matrices from different frequencies are superimposed to form the spatial-frequency feature representation of EEG signals, that is, the construction of the spatial-frequency feature representation $A^S = (A_1^S, A_2^S, \ldots, A_B^S) \in \mathbb{R}^{H\times W\times B}$ is completed.

B. EEG EMOTION RECOGNITION BASED ON BI-BRANCH ViT

Transformer is a novel neural network architecture that was primarily created for natural language processing applications, in which multi-layer perceptron layers are utilized on top of multi-head attention mechanisms to capture the long-range dependencies in sequential input. Vision Transformer has recently demonstrated considerable promise in a variety of computer vision applications, such as picture classification and segmentation [61]. Motivated by these works, we propose a new kind of ViT, the Bi-branch Vision Transformer Network, which uses a different type of EEG feature representation. Specifically, we propose a ViT architecture with two branches, each of which processes a different EEG feature representation before combining the results for EEG categorization.

We propose Bi-ViTNet to take advantage of the spatial-temporal representation and spatial- frequency representation for emotion recognition. Figure 4 illustrates the Bi-ViTNet model framework for EEG feature learning. The input of the Bi-ViTNet model is the spatial-temporal feature representation $A^T = (A_1^T, A_2^T, \ldots, A_T^T) \in \mathbb{R}^{H\times W\times T}$ and the spatial-frequency feature representation $A^S = (A_1^S, A_2^S, \ldots, A_B^S) \in \mathbb{R}^{H\times W\times B}$. We expand the spatial-temporal feature representation of size $H \times W \times T$ and the $H \times W \times B$ spatial-frequency feature representation into T spatial-temporal representation patches $A_t^T \in \mathbb{R}^{H\times W}$ with a size of $H \times W$ and B spectral-spatial representation patches $A_b^S \in \mathbb{R}^{H\times W}$ with a size of $H \times W$, respectively. Each representation patch is used as the input of the Linear Embedding layer. Linear Embedding is used to map representation patches to E_d with a constant size. According to Equation (2), W_A can be obtained as the input of the Transformer Encoder, where $x_p^{cls} \in \mathbb{R}^{E_d}$ denotes the class token in the feature representation learning, $N_{TB} \in \{T, B\}$ is the number of spatial-frequency/temporal representation patches, $E_A \in \mathbb{R}^{H\times W\times E_d}$ is the linear projection matrix, and $A_E^{pos} \in \mathbb{R}^{(N_{TB}+1)\times E_d}$ is one-dimensional position embedding, aiming to preserve the order information of frequency and time series.

$$W_A = \left[x_p^{cls}; x_p^1 E_A; x_p^2 E_A; \ldots; x_p^{N_{TB}} E_A\right] + A_E^{pos} \qquad (2)$$

As shown in Figure 4, the transformer encoder block includes multi-head self-attention, layer normalization, and

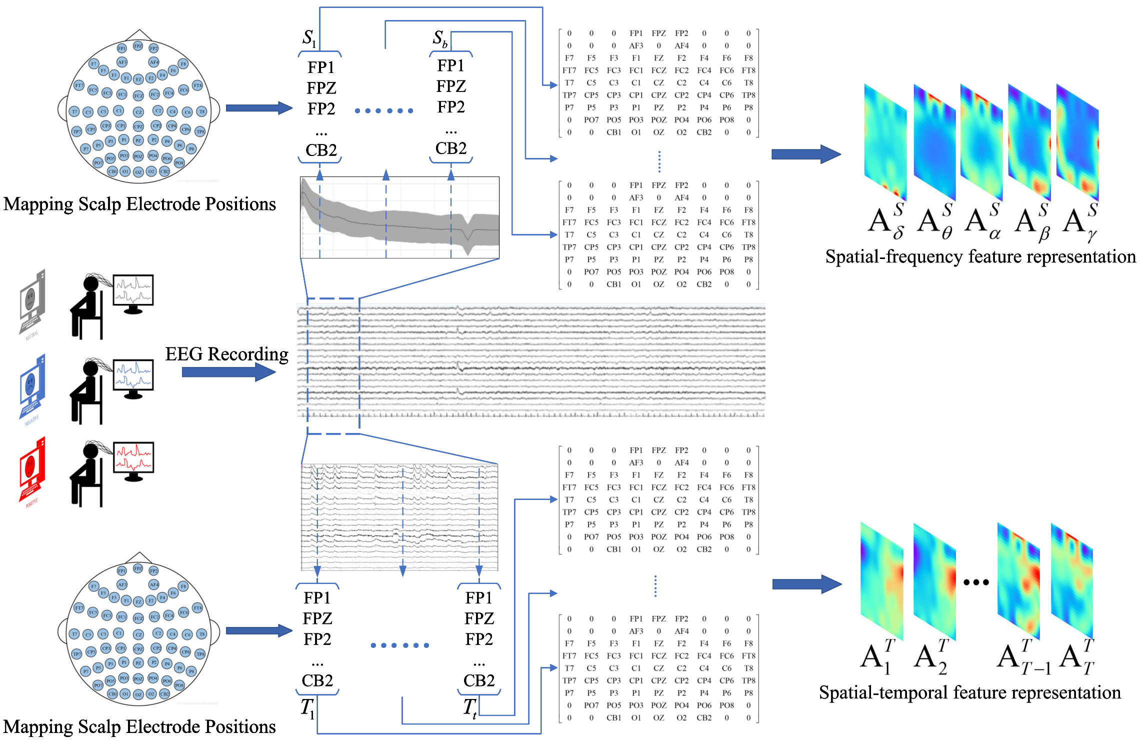

Mapping Scalp Electrode Positions
S_1
FP1
FPZ
FP2
...
CB2
S_b
EEG Recording
0 0 0 FP1 FPZ FP2 0 0 0
0 0 0 AF3 0 AF4 0 0 0
F7 F5 F3 F1 FZ F2 F4 F6 F8
FT7 FC5 FC3 FC1 FCZ FC2 FC4 FC6 FT8
T7 C5 C3 C1 CZ C2 C4 C6 T8
TP7 CP5 CP3 CP1 CPZ CP2 CP4 CP6 TP8
P7 P5 P3 P1 PZ P2 P4 P6 P8
0 PO7 PO5 PO3 POZ PO4 PO6 PO8 0
0 0 CB1 O1 OZ O2 CB2 0 0
A_δ^S A_θ^S A_α^S A_β^S A_γ^S
Spatial-frequency feature representation
T_1
T_t
A_1^T A_2^T A_{T-1}^T A_T^T
Spatial-temporal feature representation

FIGURE 3. Process of converting original EEG signal into spatial-temporal representation and spatial-frequency representation. Firstly, the original EEG signal is divided into segments with a fixed length. Secondly, the temporal-domain features of different time stamps and frequency-domain features of d ifferent frequency bands are extracted from each fragment, respectively. Finally, these features are mapped to the electrode position matrix to construct the spatial-temporal and spatial-frequency feature representations.

multiple-layer perception. The first sub-layer is the Multi-head self-attention (MSA) and the second is the Multi-layer perceptron (MLP). Before data enters each sub-layer, it is normalized by layer normalization (LN), and after it passes through each sub-layer, it is fused directly with the input using a residual connection. The operation in the transformer encoder is shown in Formula (3) and Formula (4).

$$W_l' = MSA(LN(W_{l-1})) + W_{l-1}, l = 1, \ldots, L \tag{3}$$

$$W_l = MLP(LN(W_l')) + W_l', l = 1, \ldots, L \tag{4}$$

where $MSA(\cdot)$ and $MLP(\cdot)$ represent the MSA operation and MLP operation, W_l' and W_l are the outputs of the MSA and MLP, respectively. L is the number of stacked transformer encoder blocks. Finally, the spatial-temporal feature and spatial- frequency feature representations output by the transformer encoder are sent to the Fusion and Classification layer for emotion classification based on EEG.

C. FUSION AND CLASSIFICATION

Taking the spatial-temporal features representation and spatial-frequency features representation as the input, the Bi-branch Transformer model used for spatial-frequency-temporal features fusion and classification extracts the spatial-temporal feature information and spatial-frequency feature information from the spatial-temporal feature extraction module and the spatial-frequency feature extraction module, respectively. Finally, according to Equation (5), the output from the Bi-branch transformer is fused in the Fusion and Classification layer for high-precision classification.

$$Y_{out} = \text{Softmax}(X_1^S \parallel X_2^T) \tag{5}$$

where $\parallel$ represents the concatenate operation,X_1^S and X_2^S denote the outputs from Bi-branch transformer, Y_{out} denotes the classification result of Bi-ViTNet.The cross-entropy loss is used as a loss function in this paper.

V. EXPERIMENTS

In this section, we first describe the datasets used in the study. Next, the experiment setup is then described. Finally, the experiment results are presented and discussed.

A. DATASETS

The study was carried out using SEED [53] datasets and SEED-IV [66] datasets. SEED datasets are public EEG datasets mainly used for emotion recognition. There are EEG data of 15 subjects in the datasets. Specifically, 15 Chinese film clips were selected to stimulate the subjects. Each clip viewing process can be divided into four stages, including 5s start prompt, 4-minute clip period, 45s self-assessment,

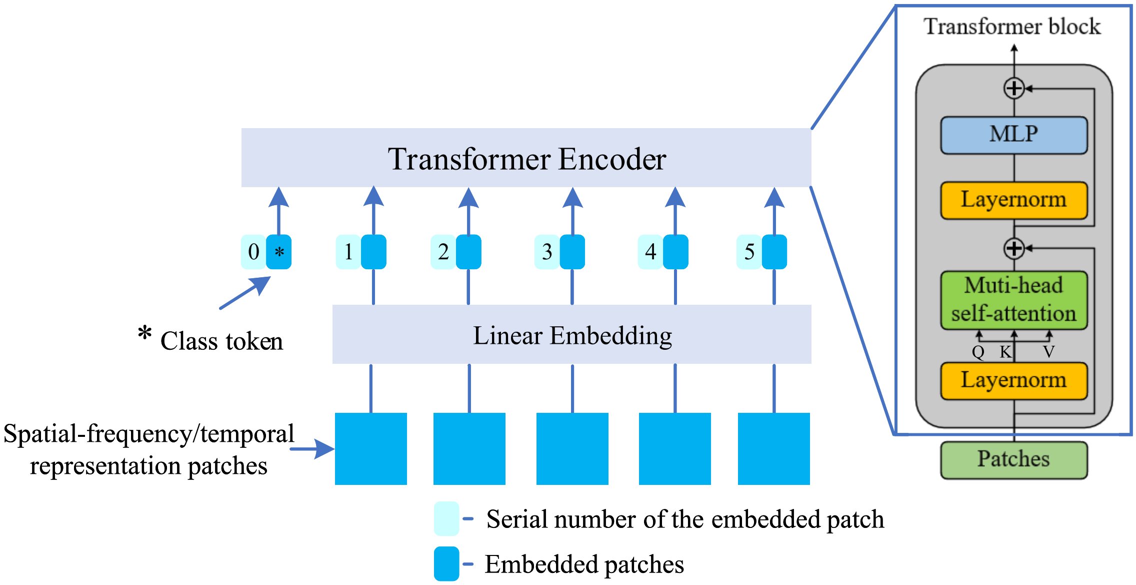


FIGURE 4. **Branch frame of Bi-ViTNet model for EEG feature learning. The Bi-ViTNet model is composed of two branches: spatial-frequency and spatial-temporal. The two branches have the same structure, which is composed of a Linear Embedding layer and a Transformer Encoder layer.**

and 15s rest period. The EEG recordings were carried out three times on each subject, and the interval between two consecutive recordings is two weeks. In every session, each subject watched 15 movie clips, each clip was about 4 minutes long, and they evoked positive, neutral, and negative emotions respectively. SEED-IV datasets are an extension of SEED datasets, including four different types of emotions, and the datasets can also be used to evaluate EEG based emotion recognition models. During the experiment, 72 video clips with a duration of 2 minutes were used to stimulate subjects to evoke happy, sad, fear, and neutral emotions.The participants self-evaluation their emotion after the video ends. ESI Neuroscan system was used to record signals of 62 channel EEG with a sampling rate of 1000 Hz, which was down sampled to 200 Hz. In order to filter noise and eliminate artifacts, EEG data of the two data sets were input to a band-pass filter. And the Power Spectral Density and Differential Entropy features of each segment in five frequency bands (δ:1 $\sim$ 4Hz, θ:4 $\sim$ 8Hz, α:8 $\sim$ 14Hz, β:14 $\sim$ 31Hz, γ: 31 $\sim$ 50Hz) were extracted. Table 1 summarizes the processing done to extract the EEG data.

B. SETTINGS

We trained the Bi-ViTNet model using a Tesla V100-SXM2-32GB GPU. A total of 12 transformer blocks were used. Each block consists of 12 attention heads. Adam optimizer was used for the training. In each experiment, we divided the original EEG signals into non-overlapping periods lasting for $\tau = 1$ seconds. Each original EEG signal can be divided into approximately 256 segments. We randomly shuffled the samples. The data was divided for training and test. The ratio between training set and test set is 7:3. The hyperparameters were as follows:

- patch_size - 32 - Size of each patch.
- num_classes - 3 or 4 - Number of classes to classify. SEED datasets are 3. SEED-IV datasets are 4.
- dim - 768 - Last dimension of output tensor after linear transformation.
- mlp_dim - 3072 - Dimension of the MLP (FeedForward) layer.
- depths - 12 - Number of Transformer blocks.
- heads - 12 - Number of heads in Multi-head Attention layer.
- dropout - 0.2 - Dropout rate.

TABLE 1. **The processing done to extract the SEED and SEED-IV datasets.**

	SEED	SEED-IV
Number of electrodes	62	62
Type of emotion	Positive Neutral Negative	Happy Sad Fear Neutral
Number of video clips	15	24
Sampling rate	200HZ	200HZ
Number of subjects	15	15
Bandpass frequency filter	0 $\sim$ 75HZ	1 $\sim$ 75HZ
Number of sessions	3	3
Frequency band	δ:1 $\sim$ 4Hz θ:4 $\sim$ 8Hz α:8 $\sim$ 14Hz β:14 $\sim$ 31Hz γ: 31 $\sim$ 50Hz	δ:1 $\sim$ 4Hz θ:4 $\sim$ 8Hz α:8 $\sim$ 14Hz β:14 $\sim$ 31Hz γ:31 $\sim$ 50Hz

C. BASELINE MODELS

We compared the proposed Bi-ViTNet with other competitive models.

- SVM [36]: A classifiers based on a least-squares support vector machine.
- DBN [53]: Deep belief networks that were trained with differential entropy features taken from multichannel EEG data to study crucial frequency bands and channels.
- DGCNN [67]: Multi-channel EEG emotion recognition using convolutional dynamical graph networks.
- RGNN [68]: Regularized graph neural network that takes the biological architecture of different brain areas into account in order to capture both global and local relationships between different EEG channels.
- R2G-STNN [30]: This method includes spatial and temporal neural network models with a regional to global hierarchical feature learning process to learn the discriminative spatial-temporal EEG features.
- BiHDM [69]: Bi-hemispheric discrepancy model that considers asymmetry discrepancies between the two hemispheres for EEG emotion identification and use four directed RNNs to obtain a deep representation of all the electrodes of EEG signals.
- SST-EmotionNet [39]: SST-EmotionNet extracts spatial, spectral, and temporal features using a two-stream network. In addition, SST-EmotionNet uses attention mechanisms to increase its EEG emotion recognition ability.
- 4D-aNN [59]: This method uses four-dimensional attention-based neural network with 4D spatial-spectral-temporal representations for EEG emotion recognition.

D. EVALUATION METRICS

For the proposed Bi-ViTNet method, its performance will be evaluated based on the following metrics: average accuracy (ACC) and standard deviation (STD). The accuracy rate is defined as the ratio of correctly identified positive and negative samples to the total number of samples, as shown in Equation (6):

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where *TP* represents the number of predicted positive samples in the positive samples, *TN* represents the number of predicted negative samples in the negative samples, *FP* represents the number of predicted positive samples in the negative samples, and *FN* represents the number of predicted negative samples in the positive samples. The standard deviation is shown in the Equation (7):

$$\mathrm{STD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{7}$$

E. RESULTS ANALYSIS AND COMPARISON

We compare Bi-ViTNet model with the baseline models using SEED and SEED-IV datasets.We evaluate the performance of the models using the accuracy and the standard deviation. Table 2 shows the average accuracy and the standard deviation of these EEG based emotion recognition models on the SEED datasets. The proposed Bi-ViTNet achieved better performance compared to the baseline models on the SEED datasets. The result shows that the performance of deep learning models were better than that of SVM classifier. DGCNN only models evaluates the spatial information of EEG data obtained from several channels and extracts the spatial information using graph convolution. BiHDM uses bidirectional-RNN to model spatial information of EEG signals, and the classification accuracy was 93.12%. R2G-STNN not only extracts the EEG electrode associations in brain regions and brain regions in order to acquire spatial information, but it also extracts the dynamic information of EEG signals in order to obtain temporal information with good accuracy. SST-EmotionNet comprehensively considers the complementarity of spatial, spectral, and temporal information, and achieved good performance, with an accuracy rate of 96.02%. 4D-aNN takes 4D spatial spectral temporal representation containing spatial, spectral, and temporal information of EEG signal as input, and integrates attention mechanism into CNN module and bidirectional LSTM module, with an accuracy rate of 96.25%. Bi-ViTNet not only considers spatial, frequency, and temporal information but also better captures the global information of EEG signals, which enables Bi-ViTNet to fully extract valuable features from EEG signals for emotion recognition. Compared with the baseline models, the accuracy of Bi-ViTNet was significantly higher. In addition, Figure 5 shows the confusion matrix of Bi-ViTNet on the SEED datasets. The results show that for Bi-ViTNet, neutral emotions are easier to identify than negative emotions and positive emotions [22].

TABLE 2. The performance on the SEED datasets.

Model	ACC(%)	STD(%)
SVM	83.99	9.72
DBN	86.08	8.34
DGCNN	90.40	8.49
BiHDM	93.12	6.06
R2G-STNN	93.38	5.96
RGNN	94.24	5.95
SST-EmotionNet	96.02	2.17
4D-aNN	96.25	1.86
Bi-ViTNet	97.55	1.58

Table 3 shows the performance of all models on the SEED-IV dataset. The proposed Bi-ViTNet achieves the state-of-the-art performance on the SEED-IV dataset. For four categories of classification tasks, the accuracy rate of DBN is 66.77%, and the accuracy rate of DGCNN and RGNN based on graph is further improved by 69.88% and 79.37% respectively. BiHDM makes full use of the difference between the two hemispheres recorded by EEG, reaching 74.35%. SST-EmotionNet and 4D-aNN add attention mechanism to learn emotional features in different fields, reaching 84.92% and 86.77% respectively. Compare with the baseline model, Bi-ViTNet has further improved the accuracy of the model to 88.08%. In addition, the confusion matrix in Figure 6 shows that Bi-ViTNet has a good recognition effect

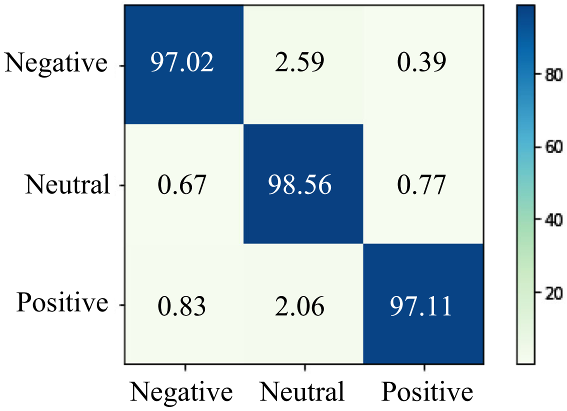


FIGURE 5. The Confusion Matrix of SEED datasets.

TABLE 3. The performance on the SEED-IV datasets.

Model	ACC(%)	STD(%)
SVM	56.61	20.05
DBN	66.77	7.38
DGCNN	69.88	16.29
BiHDM	74.35	14.09
R2G-STNN	68.51	12.37
RGNN	79.37	10.54
SST-EmotionNet	84.92	6.66
4D-aNN	86.77	7.29
Bi-ViTNet	88.08	6.32

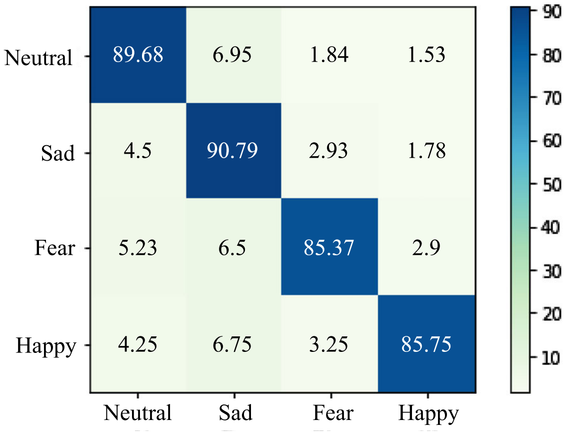


FIGURE 6. The Confusion Matrix of SEED-IV datasets.

on sad and neutral emotions, and the recognition accuracy of fear and happy is similar.

F. ABLATION STUDIES

In order to evaluate the contribution of each component of Bi-ViTNet, we conducted an ablation study. We constructed two models: a spatial-temporal encoder with a classification layer, a spatial-frequency encoder with a classification layer. Figure 7 compares the accuracies of the three models on SEED and SEED-IV datasets, the performance of Bi-ViTNet is better than that of the single encoder models. By combinging the two encoders, the accuracy improved by 6.85% and 24.71% on the SEED datasets and 6.24% and 28.38% on the SEED-IV datasets, respectively, compared with that of the spatial-frequency branch model and the spatial-temporal branch model. The results show that the two branch structures effectively use the spatial-frequency-temporal features, and the different features are complementary, which improves the classification accuracy. In addition, the branching model considering only spatial-frequency features has better performance than the branching model considering only spatial-temporal features. This shows that the importance of different features is different.

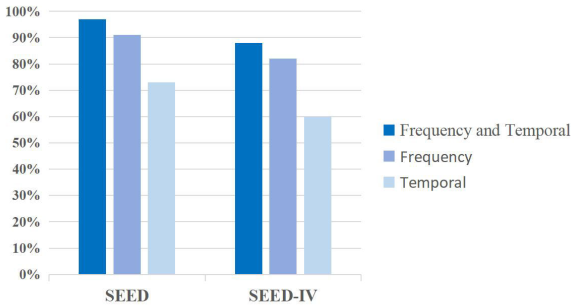


FIGURE 7. Ablation studies on Bi-branch fusion.

VI. CONCLUSION

In this paper, we propose the Bi-ViTNet, a Bi-banch Vision Transformer-based model for emotion recognition of EEG signals. The frequency features and temporal features are mapped into the electrode position matrix to construct the spatial-temporal feature representations and the spatial-frequency feature representations. Bi-ViTNet then effectively utilizes the complementarity between different features by using spatial-frequency feature representations and spatial-temporal feature representations as input. In addition, Transformer Encoder in each branch of Bi-ViTNet can better capture the global information of EEG signals. Experiments on the SEED and SEED-IV datasets show that the Bi-ViTNet model outperform all baselines. In addition, the ablation studies show the effectiveness of dual branching and the fusion of spatial-frequency-temporal features in the model. The Bi-ViTNet could also be applied to other areas, such as driving fatigue analysis and motion imagery classification. While the transformer-based model has shown excellent recognition performance, the model has a large number of parameters. In the future, we will study the lightweight transformer models for emotion recognition.

REFERENCES

[1] X. Zuo, C. Zhang, T. Hämäläinen, H. Gao, Y. Fu, and F. Cong, "Cross-subject emotion recognition using fused entropy features of EEG," *Entropy*, vol. 24, no. 9, p. 1281, Sep. 2022.

[2] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, "HetEmotionNet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1047–1056.

[3] C. Huang, Z. Han, M. Li, X. Wang, and W. Zhao, "Sentiment evolution with interaction levels in blended learning environments: Using learning analytics and epistemic network analysis," *Australas. J. Educ. Technol.*, vol. 37, no. 2, pp. 81–95, May 2021.

[4] B. Wu, Z. Liu, Q. Gu, and F.-S. Tsai, "Underdog mentality, identity discrimination and access to peer-to-peer lending market: Exploring effects of digital authentication," *J. Int. Financial Markets, Institutions Money*, vol. 83, Mar. 2023, Art. no. 101714.

[5] X. Qin, Z. Liu, Y. Liu, S. Liu, B. Yang, L. Yin, M. Liu, and W. Zheng, "User OCEAN personality model construction method using a BP neural network," *Electronics*, vol. 11, no. 19, p. 3022, Sep. 2022.

[6] Z. Xiong, Q. Liu, and X. Huang, "The influence of digital educational games on preschool children's creative thinking," *Comput. Educ.*, vol. 189, Nov. 2022, Art. no. 104578.

[7] Y. Lin, H. Song, F. Ke, W. Yan, Z. Liu, and F. Cai, "Optimal caching scheme in D2D networks with multiple robot helpers," *Comput. Commun.*, vol. 181, pp. 132–142, Jan. 2022.

[8] L. Fiorini, G. Mancioppi, F. Semeraro, H. Fujita, and F. Cavallo, "Unsupervised emotional state classification through physiological parameters for social robotics applications," *Knowl.-Based Syst.*, vol. 190, Feb. 2020, Art. no. 105217.

[9] Z. Jia, Y. Lin, J. Wang, X. Ning, Y. He, R. Zhou, Y. Zhou, and L.-W.-H. Lehman, "Multi-view spatial–temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1977–1986, 2021.

[10] H. Huang, "An EEG-based brain computer interface for emotion recognition and its application in patients with disorder of consciousness," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 832–842, Oct./Dec. 2019.

[11] Z. Jia, Y. Lin, J. Wang, R. Zhou, X. Ning, Y. He, and Y. Zhao, "GraphSleepNet: Adaptive spatial–temporal graph convolutional networks for sleep stage classification," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1324–1330.

[12] W. Liu, J. Qian, Z. Yao, X. Jiao, and J. Pan, "Convolutional two-stream network using multi-facial feature fusion for driver fatigue detection," *Future Internet*, vol. 11, no. 5, p. 115, 2019.

[13] J. Xu, S. Pan, P. Z. H. Sun, S. Hyeong Park, and K. Guo, "Human-factors-in-driving-loop: Driver identification and verification via a deep learning approach using psychological behavioral data," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3383–3394, Mar. 2023.

[14] J. Xu, K. Guo, and P. Z. Sun, "Driving performance under violations of traffic rules: Novice vs. experienced drivers," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 4, pp. 908–917, Dec. 2022.

[15] F. Wang, H. Wang, X. Zhou, and R. Fu, "A driving fatigue feature detection method based on multifractal theory," *IEEE Sensors J.*, vol. 22, no. 19, pp. 19046–19059, Oct. 2022.

[16] B. Blankertz, L. Acqualagna, S. Dähne, S. Haufe, M. Schultze-Kraft, I. Sturm, M. Uščumlic, M. A. Wenzel, G. Curio, and K.-R. Müller, "The Berlin brain-computer interface: Progress beyond communication and control," *Frontiers Neuroscience*, vol. 10, p. 530, Nov. 2016.

[17] H. Wang, Z. Cui, R. Liu, L. Fang, and Y. Sha, "A multi-type transferable method for missing link prediction in heterogeneous social networks," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 2023, doi: 10.1109/TKDE.2022.3233481.

[18] T. Li, Y. Li, M. A. Hoque, T. Xia, S. Tarkoma, and P. Hui, "To what extent we repeat ourselves? Discovering daily activity patterns across mobile app usage," *IEEE Trans. Mobile Comput.*, vol. 21, no. 4, pp. 1492–1507, Apr. 2022.

[19] X. Xie, Y. Tian, and G. Wei, "Deduction of sudden rainstorm scenarios: Integrating decision makers' emotions, dynamic Bayesian network and DS evidence theory," *Natural Hazards*, pp. 1–21, Dec. 2022.

[20] F. Shen, G. Dai, G. Lin, J. Zhang, W. Kong, and H. Zeng, "EEG-based emotion recognition using 4D convolutional recurrent neural network," *Cognit. Neurodynamics*, vol. 14, no. 6, pp. 815–828, Dec. 2020.

[21] S. Xu, Q. He, S. Tao, H. Chen, Y. Chai, and W. Zheng, "Pig face recognition based on trapezoid normalized pixel difference feature and trimmed mean attention mechanism," *IEEE Trans. Instrum. Meas.*, vol. 32, 2022, Art. no. 3500713.

[22] M. Li, M. Qiu, L. Zhu, and W. Kong, "Feature hypergraph representation learning on spatial–temporal correlations for EEG emotion recognition," *Cognit. Neurodynamics*, pp. 1–11, Oct. 2022.

[23] Z. Gao, X. Wang, Y. Yang, Y. Li, K. Ma, and G. Chen, "A channel-fused dense convolutional network for EEG-based emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 4, pp. 945–954, Dec. 2021.

[24] W. Lei, Z. Hui, L. Xiang, Z. Zelin, X. Xu-Hui, and S. Evans, "Optimal remanufacturing service resource allocation for generalized growth of retired mechanical products: Maximizing matching efficiency," *IEEE Access*, vol. 9, pp. 89655–89674, 2021.

[25] X. Lai, B. Yang, B. Ma, M. Liu, Z. Yin, L. Yin, and W. Zheng, "An improved stereo matching algorithm based on joint similarity measure and adaptive weights," *Appl. Sci.*, vol. 13, no. 1, p. 514, Dec. 2022.

[26] S. Lu, Y. Ban, X. Zhang, B. Yang, S. Liu, L. Yin, and W. Zheng, "Adaptive control of time delay teleoperation system with uncertain dynamics," *Frontiers Neurorobotics*, vol. 16, p. 152, Jul. 2022.

[27] H. Zhu, M. Xue, Y. Wang, G. Yuan, and X. Li, "Fast visual tracking with Siamese oriented region proposal network," *IEEE Signal Process. Lett.*, vol. 29, pp. 1437–1441, 2022.

[28] R. Li, Y. Wang, and B.-L. Lu, "A multi-domain adaptive graph convolutional network for EEG-based emotion recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5565–5573.

[29] Z. Jia, X. Cai, G. Zheng, J. Wang, and Y. Lin, "SleepPrintNet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging," *IEEE Trans. Artif. Intell.*, vol. 1, no. 3, pp. 248–257, Dec. 2020.

[30] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial–temporal neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 568–578, Apr. 2022.

[31] L. I. Goldfischer, "Autocorrelation function and power spectral density of laser-produced speckle patterns," *J. Opt. Soc. Amer.*, vol. 55, no. 3, pp. 247–253, Mar. 1965.

[32] D. Bansal and R. Mahajan, *EEG-Based Brain-Computer Interfaces: Cognitive Analysis and Control Applications*. New York, NY, USA: Academic, 2019.

[33] L.-H. Guo, S. Cheng, J. Liu, Y. Wang, Y. Cai, and X.-C. Hong, "Does social perception data express the spatio-temporal pattern of perceived urban noise? A case study based on 3,137 noise complaints in fuzhou, China," *Appl. Acoust.*, vol. 201, Dec. 2022, Art. no. 109129.

[34] X. Zhang, D. Huang, H. Li, Y. Zhang, Y. Xia, and J. Liu, "Self-training maximum classifier discrepancy for EEG emotion recognition," *CAAI Trans. Intell. Technol.*, pp. 1–12, Feb. 2023.

[35] Z. Wang, Z. Zhou, H. Shen, Q. Xu, and K. Huang, "JDAT: Joint-dimension-aware transformer with strong flexibility for EEG emotion recognition," *TechRxiv*, 2021.

[36] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[37] M. Li, H. Xu, X. Liu, and S. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification," *Technol. Health Care*, vol. 26, no. S1, pp. 509–519, Jul. 2018.

[38] K.-E. Ko, H.-C. Yang, and K.-B. Sim, "Emotion recognition using EEG signals with relative power values and Bayesian network," *Int. J. Control, Autom. Syst.*, vol. 7, no. 5, pp. 865–870, Oct. 2009.

[39] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "SST-EmotionNet: Spatial-spectral-temporal based attention 3D dense network for EEG emotion recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2909–2917.

[40] Y. Wang, Z. Huang, B. McCane, and P. Neo, "EmotioNet: A 3-D convolutional neural network for EEG-based emotion recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2018, pp. 1–7.

[41] Y. Shi, X. Xu, J. Xi, X. Hu, D. Hu, and K. Xu, "Learning to detect 3D symmetry from single-view RGB-D images with weak supervision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4882–4896, Apr. 2023.

[42] J. Zhang, Y. Tang, H. Wang, and K. Xu, "ASRO-DIO: Active subspace random optimization based depth inertial odometry," *IEEE Trans. Robot.*, vol. 39, no. 2, pp. 1496–1508, Apr. 2023.

[43] S. Hwang, K. Hong, G. Son, and H. Byun, "Learning CNN features from DE features for EEG-based emotion recognition," *Pattern Anal. Appl.*, vol. 23, pp. 1323–1335, Dec. 2019.

[44] S. Alhagry, A. Aly, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 1–4, 2017.

[45] Z. Jia, Y. Lin, Y. Liu, Z. Jiao, and J. Wang, "Refined nonuniform embedding for coupling detection in multivariate time series," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 101, no. 6, Jun. 2020, Art. no. 062113.

[46] Z. Jia, Y. Lin, Z. Jiao, Y. Ma, and J. Wang, "Detecting causality in multivariate time series via non-uniform embedding," *Entropy*, vol. 21, no. 12, p. 1233, Dec. 2019.

[47] Z. Jia, X. Cai, Y. Hu, J. Ji, and Z. Jiao, "Delay propagation network in air transport systems based on refined nonlinear Granger causality," *Transportmetrica B, Transp. Dyn.*, vol. 10, no. 1, pp. 586–598, Dec. 2022.

[48] Y. Liu and Z. Jia, "BSTT: A Bayesian spatial–temporal transformer for sleep staging," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–12.

[49] Z. Jia, J. Ji, X. Zhou, and Y. Zhou, "Hybrid spiking neural network for sleep electroencephalogram signals," *Sci. China Inf. Sci.*, vol. 65, no. 4, Apr. 2022, Art. no. 140403.

[50] Z. Jia, X. Cai, and Z. Jiao, "Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging," *IEEE Sensors J.*, vol. 22, no. 4, pp. 3464–3471, Feb. 2022.

[51] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, "EEG-based emotion recognition during watching movies," in *Proc. 5th Int. IEEE/EMBS Conf. Neural Eng.*, Apr. 2011, pp. 667–670.

[52] V. H. Anh, M. N. Van, B. B. Ha, and T. H. Quyet, "A real-time model based support vector machine for emotion recognition through EEG," in *Proc. Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Nov. 2012, pp. 191–196.

[53] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[54] A. Al-Nafjan, M. Hosny, A. Al-Wabil, and Y. Al-Ohali, "Classification of human emotions from electroencephalogram (EEG) signal using deep neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 9, pp. 419–425, 2017.

[55] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl. Soft Comput.*, vol. 100, Mar. 2021, Art. no. 106954.

[56] S. Liu, X. Wang, L. Zhao, J. Zhao, Q. Xin, and S.-H. Wang, "Subject-independent emotion recognition of EEG signals based on dynamic empirical convolutional neural network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 5, pp. 1710–1721, Sep. 2021.

[57] M. A. Rahman, A. Anjum, M. M. H. Milu, F. Khanam, M. S. Uddin, and M. N. Mollah, "Emotion recognition from EEG-based relative power spectral topography using convolutional neural network," *Array*, vol. 11, Sep. 2021, Art. no. 100072.

[58] A. Topic and M. Russo, "Emotion recognition based on EEG feature maps through deep learning network," *Eng. Sci. Technol., Int. J.*, vol. 24, no. 6, pp. 1442–1454, Dec. 2021.

[59] G. Xiao, M. Shi, M. Ye, B. Xu, Z. Chen, and Q. Ren, "4D attention-based neural network for EEG emotion recognition," *Cognit. Neurodynamics*, vol. 16, pp. 1–14, Jan. 2022.

[60] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, "A survey of natural language generation," *ACM Comput. Surveys*, vol. 55, no. 8, pp. 1–38, 2022.

[61] R. Hussein, S. Lee, and R. Ward, "Multi-channel vision transformer for epileptic seizure prediction," *Biomedicines*, vol. 10, no. 7, p. 1551, Jun. 2022.

[62] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[63] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[64] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "MaX-DeepLab: End-to-end panoptic segmentation with mask transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5463–5474.

[65] Z. Wang, Y. Wang, C. Hu, Z. Yin, and Y. Song, "Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4359–4368, Mar. 2022.

[66] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2018.

[67] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul./Sep. 2020.

[68] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul. 2022.

[69] Y. Yang, Q. M. J. Wu, W.-L. Zheng, and B.-L. Lu, "EEG-based emotion recognition using hierarchical network with subnetwork nodes," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 2, pp. 408–419, Jun. 2017.

WEI LU received the M.S. degree from the School of Automation and Electrical Engineering, Dalian Jiaotong University, China. He is currently pursuing the Ph.D. degree with the School of Computer Sciences, Universiti Sains Malaysia. His current research interests include affective computing, deep learning, and time series classification.

TIEN-PING TAN received the Ph.D. degree from Université Joseph Fourier, France, in 2008. He is currently a Senior Lecturer with the School of Computer Sciences, Universiti Sains Malaysia. His research interests include automatic speech recognition, machine translation, and natural language processing.

HUA MA was born in Xuchang, Henan, China, in 1977. He received the bachelor's degree from the Nanjing University of Science and Technology, China, in 2000, the master's degree from Le Mans University, France, in 2007, and the Ph.D. degree in mechanics from the University of Besancon, France, in 2011. He is currently an Assistant Professor with the Henan Province High-Speed Railway Operation and Maintenance Engineering Research Center. His research interests include signal processing, fault diagnosis, and deep learning.