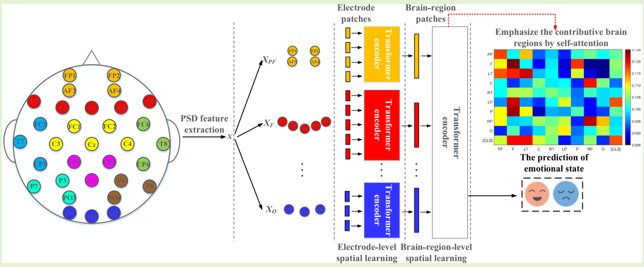
# Transformers for EEG-Based Emotion Recognition: A Hierarchical Spatial Information Learning Model

Zhe Wang, Yongxiong Wang, Chuanfei Hu, Zhong Yin, and Yu Song, *Member, IEEE*



***Abstract*—The spatial information of Electroencephalography (EEG) is essential for emotion recognition model to learn discriminative feature. The convolutional networks and recurrent networks are the conventional choices to learn the complex spatial dependencies through a number of electrodes and brain regions. However, these models have difficulty in capturing long-range dependencies due to the operations of local feature learning. To enhance EEG spatial dependencies capturing and improve the accuracy of emotion recognition, we propose a transformer- based model to hierarchically learn the discriminative spatial information from electrode level to brain-region-level. In the electrode-level spatial learning, the transformer encoders are adopted to integrate information within different brain regions. Next, in view of the different roles of brain regions in the emotion recognition, the self-attention within the transformer could emphasize the contributive brain regions. Hence, in the brain-region-level spatial learning, a transformer encoder is utilized to capture the spatial dependencies among the brain regions. Finally, to validate the effectiveness of the proposed model, the subject-independent experiments are conducted on the DEAP and MAHNOB-HCI database. The experimental results demonstrate that the proposed model achieves outstanding performance in emotion recognition with arousal and valence level. Moreover, the visualization of self-attention indicates that the proposed model could emphasize the discriminative spatial information from pre-frontal lobe, frontal lobe, temporal lobe and parietal lobe.**

***Index Terms*— Emotion recognition, EEG, transformer, self-attention.**

## I. Introduction

WITH the development of non-invasive sensor technology, the applications of physiological signals have attracted much attention [1]. Affective computing, which aims to build a harmonious human-computer interaction by endowing computers with the ability to recognize and comprehend human emotions [2], is an essential research direction among these applications. EEG [3]–[5], electromyography [6], and electro-cardiography [7] are the typical choices of physiological signals for emotion recognition. Among these modalities, EEG signals could record the activities of amygdala which is closely related to the emotions [8]. Hence, EEG-based emotion recognition has become a hot research topic in the fields of neuroscience and computer science.

In the machine learning based EEG emotion recognition methods, temporal and spectral features are commonly utilized to classify the emotions. Wang *et al.* [9] extract nonlinear dynamic EEG features from time domain, such as approximate entropy and Hurst exponent. Kroupi *et al.* [10] extract Power Spectral Density (PSD) and Wasserstein distance from different frequency bands to identify emotions. Khare *et al.* [11] propose smoothed pseudo-Wigner–Ville distribution to generate the time-frequency representation which is learned by Convolutional Neural Network (CNN). These studies have made great progress to advance EEG-based emotion recognition, but many researchers do not take full advantage of the spatial dependencies among the electrodes or brain regions, denoted as the spatial information of EEG. The research of neuroscience indicates that the human emotion is related to the pre-frontal lobe [12] and parietal lobe [13]

Manuscript received December 23, 2021; accepted January 13, 2022. Date of publication January 18, 2022; date of current version February 28, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62103299. The associate editor coordinating the review of this article and approving it for publication was Prof. Kea-Tiong (Samuel) Tang. *(Corresponding author: Yongxiong Wang.)*

Zhe Wang, Yongxiong Wang, and Zhong Yin are with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: 201440049@st.usst.edu.cn; wyxiong@usst.edu.cn; yinzhong@usst.edu.cn).

Chuanfei Hu is with the Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 214135, China (e-mail: cfhu@seu.edu.cn).

Yu Song is with the Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems, Tianjin University of Technology, Tianjin 300384, China (e-mail: jasonsongrain@hotmail.com).

Digital Object Identifier 10.1109/JSEN.2022.3144317

where these brain regions could provide more contributive information than the others. What's more, Sangnark *et al.* [14] and Lakhan *et al.* [5] find the contributive brain regions or electrodes could achieve better performance than the others by analyzing the EEG response to the popular music and movie clips. In summary, the spatial information of EEG is beneficial to discriminate emotional state.

In this work, we propose a transformer-based model, denoted as hierarchical spatial learning transformer (HSLT), to extract discriminative features by robustly capturing the EEG spatial dependencies from electrode level to brain-region level. The framework of HSLT include the following three parts:

(1) Division of the electrode patches. Inspired by the image spitting in the ViT [15], we treat PSD features from each electrode as the electrode patch. And we divide the electrode patches into different clusters according to the region classification of the cortex in the neuroscience.

(2) Electrode-level spatial learning. The electrode patches within different brain regions are separately inputted to the corresponding transformer encoders. It aims to integrate the critical information within the brain regions and lays a good foundation for learning the spatial information of overall brain.

(3) Brain-region-level spatial learning. The latent features from electrode-level transformer encoders are served as the brain region patches. And the patches are parallelly inputted to a transformer encoder to obtain the emotion prediction. The multi-head self-attention within the transformer encoder could enhance the capturing of spatial dependencies among the brain regions. Meanwhile, it could emphasize the contributive brain regions.

What's more, the learnable positional embedding and class token are adopted in all the transformer encoders of HSLT to learn essential positional information across the electrodes or brain regions, and aggregate representative information respectively.

The remainder of the paper is organized as follows. In section II, we introduce the related works. In section III, we describe the details of DEAP and MAHNOB-HCI database. In section IV, we specify the details of the HSLT for emotion recognition. In section V, we conduct extensive experiments to validate the effectiveness of HSLT. In section VI, we discuss the performance on different frequency bands, analyze the essential brain regions by performance evaluation and data visualization, compare the performance of HSLT with related works, and summarize the advantages and limitations of HSLT. Finally, we conclude this work in section VII.

## II. Related Works

### A. Spatial Information Learning for EEG-Based Emotion Recognition

Recently, many researchers apply the deep learning models to learn the spatial information of EEG. Khosrowabadi *et al.* [16] propose a biologically inspired feedforward neural network to learn the functional connectivity features between the brain regions. Although the effective spatial information could be found by this network, it is difficult for the feedforward neural network to learn the complex connectivity through a number of electrodes. Bashivan *et al.* [17] utilize CNN to robustly learn the spatial information of EEG using generated 2D EEG images according to the topology of the electrodes. Zhang *et al.* [18] propose a Recurrent Neural Network (RNN)-based model to capture the spatial dependencies among the electrodes. However, both convolutional and recurrent operation focus on the local neighborhood in space or time [19]. These will make the CNNs and RNNs have difficulty in capturing the long-range EEG spatial dependencies which is beneficial to extract critical information through a number of electrodes and brain regions.

### B. Transformer and Its Application in Different EEG Tasks

To enhance the long-range dependencies capturing in EEG, attention mechanism based method [20], [21] have been adopted for improving the emotion recognition performance. As the self-attention based deep learning model, transformer [22] has attracted more and more attention. Transformer-based models have achieved the state-of-the-art performance in the fields of NLP and CV, such as the BERT [23] and ViT [15]. The multi-head self-attention and parallel inputting make the transformers have superior abilities to capture the long-range dependencies. Besides, the positional embedding retains the critical positional information of words and image patches, and class token could aggregate representative information. These strategies could potentially improve the performance of sequence analysis [15].

Recently, transformer-based models also have been proposed for different EEG tasks. Sun *et al.* [24] combine the transformer with CNN to improve the accuracy of motor imagery. Pedoeem *et al.* [25] build a hybrid architecture of convolutional layers, fully connected layers, and a transformer for seizure detection. These researches indicate that transformer is also a powerful model for extracting discriminative EEG features.

## III. The Description of the Emotion Databases

In this work, the DEAP database [26] and MAHNOB-HCI [27] are chosen as the benchmark emotion databases. Here, we briefly introduce the experimental protocol and preprocessing of the EEG data.

The DEAP is a multimodal emotion database which includes EEG signals and peripheral signals from 32 subjects. The 40 one-minute-long music videos are utilized to elicit the emotions, and these videos are presented in 40 trails. In each trail, the EEG signals and peripheral signals are simultaneously recorded when subjects watching the videos. At the end of each trail, the subjects are rated the arousal, valence, and dominance according to Self-Assessment Manikin (SAM) with a range of 1 to 9.

The MAHNOB-HCI database is built for analysis emotions and implicit tagging. The experimental protocol of MAHNOB-HCI is similar to the DEAP. 20 movie clips are presented which are ranged from 34.9 seconds to 117 seconds long. During each trail, EEG signals and peripheral signals are recorded from 27 subjects, and subjects accomplish the self-report using arousal, valence, dominance, predictability and

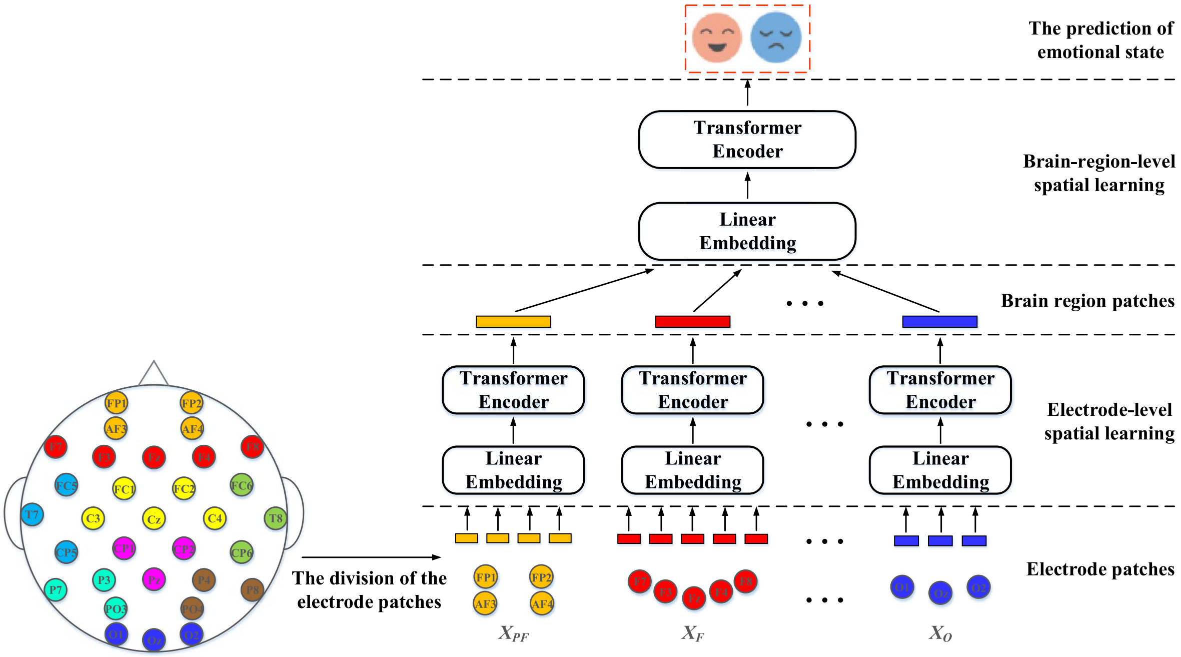


Fig. 1. The overview of the HSLT. We extract the PSD features from each electrode and treat the aggregation of PSD features from each electrode as the electrode patch. Next, electrode patches are divided according to the layout of the brain regions, the electrodes with the same color are divided into the same brain region. And the EEG spatial information learning includes two stages: the electrode-level spatial learning and the brain-region-level spatial learning.

emotional keywords. The EEG acquisition equipment is also the 32-channel Biosemi system, so the preprocessing of EEG data is consistent with the DEAP. But the EEG data of 24 subjects are available due to the incomplete data of 3 subjects.

The EEG signals of two datasets are both recorded by a 32-channel Biosemi system with a sampling rate of 512 Hz, and the electrodes are placed according to the 10-20 system. Firstly, the EEG signals are downsampled to 128 Hz. Next, a 4-45 Hz bandpass filter is applied. Finally, the electrooculography (EOG) artifacts are removed by the blind source separation technique using EEGLAB toolbox [28].

## IV. The Novel HSLT for EEG Emotion Recognition

### A. The Overview of the HSLT

In this section, we will identify the configuration of the HSLT. As shown in Fig. 1., HSLT includes three major parts: EEG feature extraction and division of the electrode patches, electrode-level spatial learning and brain-region-level spatial learning. We will briefly introduce these three parts in the remainder of this section.

### B. EEG Feature Extraction and Division of the Electrode Patches

For each EEG segment, we extract the PSD features from different frequency bands. What's more, the PSD features are obtained by the Welch method with a one-second-long Hamming window. The aggregation of PSD features from one electrode, denoted as the electrode patch, can be formulated as $x_i \in \mathbb{R}^d$, where $i$ denotes the $i$-th electrode and $d$ is number of the frequency bands. And the EEG feature set can be denoted as $X = [x_1, x_2, \ldots, x_e] \in \mathbb{R}^{e\times d}$, where $e$ is the number of electrodes.

TABLE I
The Electrode Patch Clusters Associated With Different Brain Regions

| Brain region | Electrodes | Representation of the clusters |
|---|---|---|
| Pre-frontal (PF) | FP1, AF3, AF4, FP2 | $X_{PF} \in \mathbb{R}^{4\times d}$ |
| Frontal (F) | F7, F3, Fz, F4, F8 | $X_F \in \mathbb{R}^{5\times d}$ |
| Left temporal (LT) | FC5, T7, CP5 | $X_{LT} \in \mathbb{R}^{3\times d}$ |
| Central (C) | FC1, C3, Cz, C4, FC2 | $X_C \in \mathbb{R}^{5\times d}$ |
| Right Temporal (RT) | FC6, T8, CP6 | $X_{RT} \in \mathbb{R}^{3\times d}$ |
| Left Parietal (LP) | P7, P3, PO3 | $X_{LP} \in \mathbb{R}^{3\times d}$ |
| Parietal (P) | CP1, Pz, CP2 | $X_P \in \mathbb{R}^{3\times d}$ |
| Right Parietal (RP) | P8, P4, PO4 | $X_{RP} \in \mathbb{R}^{3\times d}$ |
| Occipital (O) | O1, Oz, O2 | $X_O \in \mathbb{R}^{3\times d}$ |

The brain response to different emotions could be reflected by the activations of various brain regions [29], [30]. Hence, the differences among the activations of the brain regions could supply the informative feature to distinguish the emotions. The lobes of the brain are the anatomical classification of the cortex, the frontal lobe, central lobe, temporal lobe, parietal lobe and occipital lobe are related to different brain functions [31]. According to the criterion, we split the electrode patches into different clusters to represent the functions of various brain regions. As shown in the Fig. 1. and Table I, the pre-frontal, frontal, and so on are the nine clusters utilized in this work.

### C. The Details of the Transformer Encoder in the HSLT

The overview of the transformer encoder in the HSLT is shown in Fig. 2. Given the representation of electrode patches

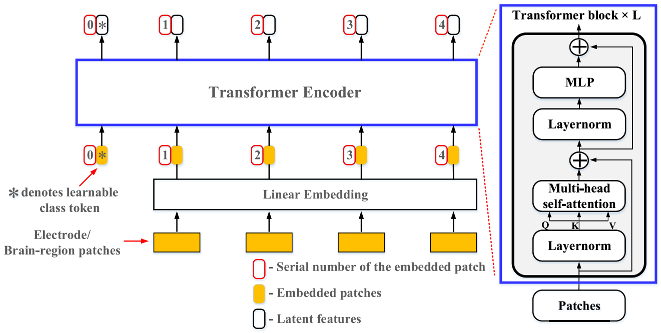


Fig. 2. The overview of the transformer encoder in the HSLT. A transformer encoder contains L stacked transformer blocks.

within the brain region $X_E = [X_E^1, X_E^2, \ldots, X_E^N] \in \mathbb{R}^{N\times d}$. Firstly, the electrode patches are mapped to a constant size $D_e$ using linear embedding. According to the Eq.1, we could obtain the representation of patch embeddings $E \in \mathbb{R}^{(N+1)\times D_e}$, where $x_E^{cls} \in \mathbb{R}^{D_e}$ denotes the class token in the electrode-level learning. The class token is an extra learnable embedding which could aggregate representative information of all embeddings [18]. Besides, $W_E \in \mathbb{R}^{d\times D_e}$ is the linear projection matrix.

$$E = \left[x_E^{cls}; x_E^1 W_E; x_E^2 W_E; \ldots; x_E^N W_E\right] \tag{1}$$

Next, we utilized the 1-D positional embedding $E_E^{pos} \in \mathbb{R}^{(N+1)\times D_e}$ which aims to retain the spatial information of the electrode patches or region patches. According to the Eq.2, the result $Z_E$ is served as input of the transformer encoder.

$$Z_E = E + E_E^{pos} \tag{2}$$

As shown in Fig.2, the transformer block includes multi-head self-attention, layer normalization and Multiple Layer Perception (MLP). The Multi-head Self-Attention (MSA) is an extension of self-attention which is widely used in NLP [32]. And the Layer Normalization (LN) could reduce the training time and improve the generalization performance [33]. Here, we given $L_e$ as the block number in the electrode-level spatial learning. The operation in the transformer encoder is shown as Eq.3 and Eq.4.

$$Z'_{l_e} = MSA\left(LN\left(Z_{l_e-1}\right)\right) + Z_{l_e-1} \quad l = 1, \ldots, L \tag{3}$$

$$Z_{l_e} = MLP\left(LN\left(Z'_{l_e}\right)\right) + Z'_{l_e} \quad l = 1, \ldots, L \tag{4}$$

where $MSA(\cdot)$ and $MLP(\cdot)$ denote MSA operation and MLP operation, $Z'_{l_e}$ and $Z_{l_e}$ are the output of the MSA and MLP respectively. The $L$ is number of the stacked transformer blocks, so the output of the last block is $Z_{L_e} \in \mathbb{R}^{(N+1)\times D_e}$.

In the electrode-level spatial learning, the electrode patches within one brain region are parallelly inputted to the corresponding transformer encoder. According to the division of electrodes in the Table I, there are nine transformer encoders which are corresponding to the brain regions. The latent features $X_L$ obtained by nine transformer encoders could be formulated as:

$$X_L = \left[Z_{L_e}^{PF}; Z_{L_e}^{F}; Z_{L_e}^{LT}; Z_{L_e}^{C}; Z_{L_e}^{RT}; Z_{L_e}^{LP}; Z_{L_e}^{P}; Z_{L_e}^{RP}; Z_{L_e}^{o}\right] \tag{5}$$

The dimensionality of each element in the $X_L$ is not equal. The reason is that the $PF$, $F$, and $C$ brain regions include more than three electrodes, while other brain regions only contain three electrodes. To guarantee the same dimensionality, $Z_{L_e}^{PF}$, $Z_{L_e}^{F}$, and $Z_{L_e}^{C}$ are projected into the $4 \times D_e$ dimensions. According to the Eq.6, we obtained the patches $X_R \in \mathbb{R}^{9\times 4\times D_e}$ for brain-region-level spatial learning.

$$\begin{aligned} X_R &= \Big[Z_{L_e}^{PF} W_{PF}; Z_{L_e}^{F} W_F; Z_{L_e}^{LT}; Z_{L_e}^{C} W_C; Z_{L_e}^{RT}; Z_{L_e}^{LP}; Z_{L_e}^{P}; \\ &\quad Z_{L_e}^{RP}; Z_{L_e}^{o}\Big] \\ &= [X_R^1, X_R^2, \ldots, X_R^9] \in \mathbb{R}^{9\times 4\times D_e} \end{aligned} \tag{6}$$

In the brain-region spatial learning, the operations are similar to the electrode-level spatial learning, and the operations can be represented as:

$$Z_R = \left[x_R^{class}; x_R^1 W_R; x_R^2 W_R; \ldots; x_R^9 W_R\right] + E_R^{pos} \tag{7}$$

$$Z'_{l_r} = MSA\left(LN\left(Z_{l_r-1}\right)\right) + Z_{l_r-1} \quad l_r = 1, \ldots, L_r \tag{8}$$

$$Z_{l_r} = MLP\left(LN\left(Z'_{l_r}\right)\right) + Z'_{l_r} \quad l_r = 1, \ldots, L_r \tag{9}$$

where the $W_R \in \mathbb{R}^{4\times D_e\times D_r}$ is the weight of the linear projection, $E_R^{pos} \in \mathbb{R}^{10\times D_r}$ is the positional embedding, $Z'_{l_r}$ and $Z_{l_r}$ are the outputs of the MSA and MLP respectively. The output $Z_{L_r}^0 \in \mathbb{R}^{D_R}$ is corresponding to the $x_R^{class}$. Finally, the prediction of arousal or valence is obtained by Eq.10.

$$\hat{y} = \sigma\left(W_O Z_{L_R}^0\right) \tag{10}$$

where $\hat{y}$ is the prediction result, $W_O$ is the weight, and $\sigma(\cdot)$ is the sigmoid function.

## V. Results

In this section, we separately conduct the subject-independent experiments on the DEAP and MAHNOB-HCI database to validate the effectiveness of HSLT. In each subject-independent experiment, the leave-one-subject-out cross-validation are adopted to derive the classification performance. Specifically, EEG data from one subject is considered as the testing set, and EEG data from the remaining subjects are set as training set until each subject's data has been set as test set once. And the final performance is the average of all the folds.

The sufficient data are necessary to provide for training transformer-based model [15]. Hence, we adopt a 6-second-long sliding window with 50% overlap to segment the EEG data. In this work, the self-assessment of 1-4 is treated as 'low class,' the low arousal (LA) and low valence (LV). And the self-assessment of 6-9 is treated as 'high class,' the high arousal (HA) and high valence (HV). According to the criterion, we transform the emotion recognition problem into the binary classification (LA vs. HA, and LV vs. HV) and the four-classes classification (LALV vs. LAHV vs. HALV vs. HAHV). What's more, to deal with the imbalance samples problem in the two databases shown in Table II and Table III, we adopt the accuracy (denoted as $P_{acc}$), weighted F1-score (denoted as $P_f$), and Cohen's kappa coefficient (denoted as $P_{ck}$) to evaluate the classification performance.

TABLE II
THE NUMBER OF SAMPLES IN THE BINARY CLASSIFICATION

| Database | LA | HA | LV | HV |
|---|---|---|---|---|
| DEAP | 7144 | 10108 | 6973 | 10374 |
| MAHNOB -HCI | 8538 | 6585 | 8588 | 6241 |

TABLE III
THE NUMBER OF SAMPLES IN THE FOUR-CLASS CLASSIFICATION

| Database | LALV | LAHV | HALV | HAHV |
|---|---|---|---|---|
| DEAP | 2413 | 2014 | 2983 | 5225 |
| MAHNOB -HCI | 3097 | 2972 | 4118 | 2151 |

TABLE IV
THE RESULTS OF HSLT IN THE BINARY CLASSIFICATION

| Database | LA vs. HA | | | LV vs. HV | | |
|---|---|---|---|---|---|---|
| | $P_{acc}$ (%) | $P_f$ (%) | $P_{ck}$ | $P_{acc}$ (%) | $P_f$ (%) | $P_{ck}$ |
| DEAP | 65.75<br>(8.51) | 64.29<br>(10.06) | 0.439<br>(0.112) | 66.51<br>(8.53) | 66.27<br>(7.29) | 0.445<br>(0.109) |
| MAHNOB -HCI | 66.20<br>(10.48) | 63.85<br>(10.23) | 0.457<br>(0.129) | 66.63<br>(6.51) | 65.33<br>(7.21) | 0.463<br>(0.101) |

*The standard deviations are in the brackets.*

TABLE V
THE RESULTS OF HSLT IN THE FOUR-CLASS CLASSIFICATION

| Database | $P_{acc}$ (%) | $P_f$ (%) | $P_{ck}$ |
|---|---|---|---|
| DEAP | 56.93<br>(8.22) | 54.29<br>(8.59) | 0.327<br>(0.094) |
| MAHNOB-HCI | 58.03<br>(9.36) | 55.32<br>(9.71) | 0.343<br>(0.107) |

TABLE VI
THE RESULTS OF DEEP NETWORKS IN THE DEAP DATABASE (%)

| | LA vs. HA | | LV vs. HV | |
|---|---|---|---|---|
| | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| The proposed HSLT | **65.75**<br>**(8.51)** | 64.29<br>(10.06) | **66.51**<br>**(8.53)** | 66.27<br>(7.29) |
| One-stage Transformer# | 61.47<br>(11.83) | 59.44<br>(11.82) | 61.89<br>(8.67) | 59.65<br>(9.25) |
| CNN* [34] | 60.72<br>(10.10) | 58.88<br>(10.83) | 61.43<br>(8.60) | 60.68<br>(8.08) |
| LSTM* [35] | 58.39<br>(9.04) | 52.61<br>(9.85) | 58.65<br>(8.16) | 53.31<br>(8.90) |
| DBN* [36] | 59.82<br>(8.72) | 57.49<br>(8.23) | 60.70<br>(8.93) | 59.72<br>(8.41) |
| CNN-LSTM* [37] | 62.10<br>(9.63) | 60.47<br>(9.99) | 63.86<br>(8.55) | 61.36<br>(7.73) |
| DenseNet* [38] | 63.37<br>(7.05) | 61.50<br>(7.47) | 64.57<br>(7.48) | 62.58<br>(7.06) |

## A. The Classification Results of HSLT

The PSD features are extracted on five bands, Theta band (4-7 Hz), Slow alpha band (8-10 Hz), Alpha band (8-12 Hz), Beta band (13-30 Hz) and Gamma band (30-47 Hz). Hence, the dimension of the feature $d$ is 5. We use the dropout with 0.4 in all the MSA modules and MLP modules. In addition, we use the dropout with 0.1 in the linear patch projections of brain-region-level learning. For training the HSLT, the Adam optimizer with cosine learning decay is adopted, batch size is set as 512, and epoch is 80 with early stopping. The loss function is cross entropy function, and labels are processed according to the one-hot coding. All the networks in this work are implemented by pytorch with two NVIDIA GeForce RTX 2080 GPUs. Additionally, we set the hyperparameter $L_e$, $L_r$ both as 2, $D_e$, $D_r$ are fixed as 8 and 16 respectively. MSA is adopted in the electrode-level and brain-region-level, $D_h$ and $k$ are set as 64 and 16 respectively.

The classification results of binary and four-class classification in two databases are listed in Table IV and Table V. In the binary classification, the HSLT achieves the accuracies of 65.75% and 66.51% with arousal and valence level in the DEAP database. Meanwhile, HSLT also obtain the accuracies of 66.20% and 66.63% in the MAHNOB-HCI database. Compared with binary classification, the performance of four-class classification properly declines due to the increasing difficulty of classification. The HSLT achieves the accuracies of 56.93% and 58.03% with arousal and valence level in the DEAP and MAHNOB-HCI database, respectively.

## B. Comparison With Other Deep Networks

The results of different methods in two databases have been illustrated in Table VI and Table VII. According to the results, the proposed HSLT has achieved optimal performance and it out- performs the other networks in both two databases. In contrast to the accuracy of CNN, LSTM and DBN, the accuracy of proposed HSLT have been boosted more than 5% and 6% in the DEAP and MAHNOB-HCI database respectively. Moreover, the proposed HSLT surpasses the one-stage Transformer over 4%. Similarly, the one-stage Transformer is also superior to these three networks. In addition, we also reimplement the latest spatial and temporal EEG encoding networks, CNN-LSTM and DenseNet. Compared with these two networks, the HSLT also achieves a slight improvement.

Except for the proposed HSLT and one-stage transformer, the attention mechanisms are not adopted in the other networks. According to the comparison, the self-attention within HSLT could enhance the long-range dependencies capturing in EEG. Besides, the standard deviation of HSLT is basically less than the other networks. It suggests that more stable performances have achieved by HSLT in the test set. In summary, these results could validate the effectiveness of the proposed HSLT in the emotion recognition.

TABLE VII
THE RESULTS OF DEEP NETWORKS IN THE MAHNOB-HCI DATABASE (%)

| | LA vs. HA | | LV vs. HV | |
|---|---|---|---|---|
| | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| The proposed HSLT | **66.20**<br>**(10.48)** | 63.85<br>(10.23) | **66.63**<br>**(6.51)** | 65.53<br>(7.21) |
| One-stage Transformer# | 61.34<br>(10.85) | 58.82<br>(10.02) | 61.51<br>(7.17) | 60.21<br>(7.20) |
| CNN* [34] | 58.15<br>(10.62) | 54.71<br>(10.54) | 59.47<br>(5.96) | 58.92<br>(6.62) |
| LSTM* [35] | 56.84<br>(10.28) | 53.87<br>(9.36) | 57.98<br>(6.89) | 53.67<br>(7.06) |
| DBN* [36] | 56.95<br>(10.93) | 53.31<br>(10.02) | 59.78<br>(9.36) | 58.59<br>(9.49) |
| CNN-LSTM* [37] | 62.31<br>(9.07) | 60.70<br>(8.53) | 62.74<br>(8.21) | 61.05<br>(8.69) |
| DenseNet* [38] | 63.85<br>(9.15) | 61.52<br>(8.76) | 64.10<br>(8.62) | 61.90<br>(7.99) |

*The model with * denotes that we conduct same experiment by our own reimplementation.*

*One-stage Transformer # means the all the 32 electrode patches are parallelly inputted into a transformer encoder. And the hyperparameters of one-stage HSLT are same to the transformer encoder of proposed HSLT in the brain-region-level spatial learning.*

TABLE VIII
THE HYPERPARAMETERS OF DIFFERENT HSLT CONFIGURATIONS

| | $L_e$ | $L_R$ | $D_e$ | $D_R$ | $D_h$ | $k$ |
|---|---|---|---|---|---|---|
| HSLT-Small | 1 | 1 | 8 | 16 | 32 | 8 |
| HSLT-Base | 1 | 2 | 8 | 16 | 32 | 12 |
| The proposed HSLT | 2 | 2 | 8 | 16 | 64 | 16 |
| HSLT-Large | 2 | 3 | 8 | 16 | 64 | 16 |

### C. HSLT With Different Configurations

In this part, we compare the performances between the different configurations of HSLT. Hence, we propose the four candidate configurations of HSLT, as the parameters summarized in Table VIII The configurations are denoted as ‘HSLT-small,’ ‘HSLT-base,’ ‘The proposed HSLT’ and ‘HSLT-Huge’ with the depth and head increasing.

The performances of different HSLT configurations in the DEAP database are listed in Table IX. The proposed HSLT achieves the best performance in the arousal and valence classification, and the performance of HSLT-Large is similar to the proposed HSLT. Furthermore, compared with HSLT-small and HSLT-base, the proposed HSLT and HSLT-Large significantly improve the accuracy with the paired t-test results ($p < 0.05$). It indicates that more discriminative feature could be learned by HSLT with the depth and head increasing. Additionally, the accuracy of proposed HSLT is significant outperformed HSLT-Large in the valence classification. And the accuracy of the proposed HSLT improve 0.13% than the HSLT-Large in the arousal classification.

TABLE IX
THE RESULTS OF DIFFERENT HSLT CONFIGURATIONS IN THE DEAP DATABASE (%)

| | LA vs. HA | | LV vs. HV | |
|---|---|---|---|---|
| | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| HSLT-Small | 63.07<br>(10.49) | 62.98<br>(8.90) | 64.03<br>(8.40) | 63.69<br>(7.75) |
| HSLT-Base | 64.29<br>(10.85) | 63.88<br>(10.02) | 64.47<br>(8.36) | 63.98<br>(7.42) |
| The proposed HSLT | **65.75**<br>**(8.51)** | 64.29<br>(10.06) | **66.51***<br>**(8.53)** | 66.27<br>(7.29) |
| HSLT-Large | 65.62<br>(9.29) | 64.18<br>(9.72) | 65.50<br>(8.85) | 64.62<br>(8.91) |

TABLE X
THE RESULTS OF DIFFERENT HSLT CONFIGURATIONS IN THE MAHNOB-HCI DATABASE (%)

| | LA vs. HA | | LV vs. HV | |
|---|---|---|---|---|
| | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| HSLT-Small | 62.79<br>(10.72) | 61.69<br>(10.37) | 64.65<br>(6.69) | 63.88<br>(8.60) |
| HSLT-Base | 64.14<br>(11.79) | 63.34<br>(12.32) | 65.56<br>(6.74) | 64.75<br>(8.23) |
| The proposed HSLT | **66.20***<br>**(10.48)** | 63.85<br>(10.23) | **66.63**<br>**(6.51)** | 65.53<br>(7.21) |
| HSLT-Large | 64.01<br>(11.61) | 62.91<br>(11.06) | 66.29<br>(7.21) | 64.82<br>(8.03) |

*The accuracy with * denotes that the method significantly outperforms other methods using paired t-test (p<0.05).*

The similar results are also shown in the MAHNOB-HCI database, as shown in Table X. The proposed HSLT and HSLT-Large are both significant outperformed other configurations in the arousal and valence classification. And the HSLT significantly outperforms HSLT-Large in the arousal classification. And the accuracy of the proposed HSLT improve 0.34% than the HSLT-Large in the valence classification. Apparently, the proposed HSLT has advantages over other configurations in the valence and arousal classification.

### D. The Analysis of the Positional Embedding and Class Token

To investigate the effect of positional embedding (PE) and class token (denoted as [CLS]), we employ the ablation experiments and results are reported in the Table XI and Table XII. It can be seen that the performance of the proposed HSLT has significantly decreased (more than 5%) when the PE or [CLS] is removed from HSLT. What’s more, the lowest accuracies have been obtained when the PE and [CLS] are both removed (decreased more than 6%). In summary, these results indicate that the PE and [CLS] are both contributed to improve the classification performance.

## VI. DISCUSSION

The outstanding results achieved by proposed HSLT have shown the in the section V. In this part, the analysis of

TABLE XI
THE RESULTS OF ABLATION EXPERIMENTS IN THE DEAP DATABASE (%)

| | LA vs. HA | | LV vs. HV | |
|---|---|---|---|---|
| | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| The proposed HSLT | **65.75***<br>**(8.51)** | 64.29<br>(10.06) | **66.51***<br>**(8.53)** | 66.27<br>(7.29) |
| Non-PE | 62.45<br>(10.74) | 61.22<br>(10.12) | 63.15<br>(9.02) | 61.84<br>(8.83) |
| Non-[CLS] | 61.76<br>(10.99) | 59.96<br>(10.86) | 62.14<br>(8.51) | 61.21<br>(7.64) |
| Non-PE and [CLS] | 61.22<br>(9.60) | 59.79<br>(9.66) | 61.69<br>(8.41) | 60.78<br>(7.86) |

TABLE XII
THE RESULTS OF ABLATION EXPERIMENTS IN THE MAHNOB-HCI DATABASE (%)

| | LA vs. HA | | LV vs. HV | |
|---|---|---|---|---|
| | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| The proposed HSLT | **66.20***<br>**(10.48)** | 63.85<br>(10.23) | **66.63***<br>**(6.51)** | 65.53<br>(7.21) |
| Non-PE | 59.67<br>(10.33) | 58.89<br>(10.07) | 60.41<br>(6.47) | 59.45<br>(6.91) |
| Non-[CLS] | 59.24<br>(9.74) | 57.49<br>(9.34) | 60.49<br>(7.07) | 59.38<br>(7.47) |
| Non-PE and [CLS] | 58.34<br>(8.86) | 58.03<br>(8.96) | 59.61<br>(5.88) | 57.99<br>(5.64) |

TABLE XIII
THE PERFORMANCE EVALUATION ON DIFFERENT FREQUENCY BANDS (%)

| Frequency | DEAP | | MAHNOB-HCI | |
|---|---|---|---|---|
| bands | LA vs. HA | LV vs. HV | LA vs. HA | LV vs. HV |
| Theta | 58.69 | 58.67 | 59.03 | 58.45 |
| Alpha | 59.64 | 60.19 | 59.39 | 61.07 |
| Slow alpha | 59.73 | 61.32 | 59.51 | 61.15 |
| Beta | 59.98 | 61.98 | 60.72 | 61.77 |
| Gamma | 60.19 | 62.03 | 60.97 | 62.54 |
| **All bands** | **65.75*** | **66.51*** | **66.20*** | **66.63*** |

the frequency bands, analysis of brain regions, comparison with related researches are conducted to further verify the effectiveness of proposed HSLT.

## *A. The Analysis of the Frequency Bands*

In this part, we conduct the performance evaluation to investigate relationship among frequency bands, and the results are shown in Table XIII. According to the results, the performance on Beta bands and Gamma bands is higher than the other bands in both arousal and valence classification. It indicates that the features from these two bands are more contributive to discriminate the emotional state. Moreover, the HSLT obtain optimal performance when using all bands features ($p < 0.05$). It suggests that the discriminative information is distributed in all the bands, and feature from different bands are complementary.

TABLE XIV
THE PERFORMANCE EVALUATION ON DIFFERENT BRAIN REGIONS (%)

| Brain | DEAP | | MAHNOB-HCI | |
|---|---|---|---|---|
| regions | LA vs. HA | LV vs. HV | LA vs. HA | LV vs. HV |
| PF | 61.25 | 61.58 | 61.96 | 62.13 |
| F | 61.04 | 61.71 | 60.91 | 62.41 |
| LT | 60.08 | 61.53 | 59.06 | 61.06 |
| C | 59.36 | 60.03 | 60.11 | 59.86 |
| RT | 60.29 | 61.02 | 59.34 | 61.92 |
| LP | 60.60 | 59.20 | 60.73 | 58.88 |
| P | 61.07 | 59.84 | 61.32 | 58.40 |
| RP | 60.76 | 60.61 | 62.07 | 58.06 |
| O | 59.75 | 60.90 | 58.38 | 58.26 |
| **All regions** | **65.75*** | **66.51*** | **66.20*** | **66.63*** |

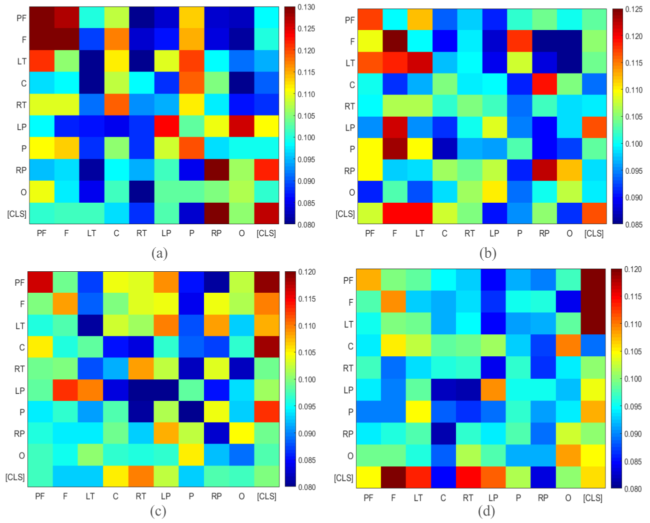


Fig. 3. The average score of the heads in the brain-region-level (a) subject25's arousal classification in the DEAP database (b) subject17's valence classification in the DEAP database (c) subject11's arousal classification in the MAHNOB-HCI database (d) subject6's valence classification in the MAHNOB-HCI database.

## *B. The Analysis of the Brain Regions*

To analysis the essential brain regions, we firstly conduct the performance evaluation on each brain region, the results are shown in Table XIV. These results are obtained by the one-stage Transformer, because the HSLT is designed for capturing the dependencies among the multiple brain regions. According to the results in the Table XIV, the PF, F, and P achieve better performance on arousal classification in the DEAP database. Meanwhile, the PF, P and RP also achieves better performance in the MAHNOB-HCI database. It indicates that the pre-frontal lobe and parietal lobe are more contributive in the arousal classification. On the other hand, the PF, F, LT and RT achieve better performance on valence classification in both DEAP and MAHNOB-HCI database. It suggests that the pre-frontal lobe, frontal lobe and temporal lobe are more contributive in the valence classification.

TABLE XV
THE DETAILS OF THE RELATED RESEARCHES USING SUBJECT-INDEPENDENT SCHEME

| Related researches | Database | Number of classes & classifier | Validation method | $P_{acc}$ (%) |
|---|---|---|---|---|
| Li *et al.* [40] (2018) | DEAP | 2-classes, SVM | Leave-one-subject-out | Arousal: -<br>Valence:59.06 |
| Pandey *et al.* [41] (2019) | DEAP | 2-classes, Deep neural network | 30 subjects for training, 2 subjects for testing | Arousal:61.25<br>Valence:62.50 |
| Rayatdoost *et al.* [39] (2018) | DEAP/ MAHNOB-HCI | 2-classes Random Forest | Leave-one-subject-out | Arousal:59.22<br>Valence:55.70<br>(DEAP)<br>Arousal:71.25<br>Valence:61.46<br>(MAHNOB-HCI) |
| Yin *et al.* [42] (2020) | DEAP/ MAHNOB-HCI | 2-classes, LSSVM | Leave-one-subject-out | Arousal:65.10<br>Valence:67.97<br>(DEAP)<br>Arousal:67.43<br>Valence:70.90<br>(MAHNOB-HCI) |
| Zhang *et al.* [43] (2020) | DEAP/ MAHNOB-HCI | 2-classes, KNN | Leave-one-subject-out | Arousal:65.21<br>Valence:66.35<br>(DEAP)<br>Arousal:65.20<br>Valence:65.37<br>(MAHNOB-HCI) |
| Huang *et al.* [44] (2016) | MAHNOB-HCI | 3-classes, SVM | Leave-one-subject-out | Arousal: 62.00<br>Valence: 61.00 |
| Soleymani *et al.* [27] (2012) | MAHNOB-HCI | 3-classes, SVM | Leave-one-subject-out | Arousal: 52.40<br>Valence: 57.00 |
| The proposed HSLT | DEAP/ MAHNOB-HCI | 2-classes, HSLT | Leave-one-subject-out | Arousal: 65.75<br>Valence: 66.51<br>(DEAP)<br>Arousal:66.20<br>Valence:66.63<br>(MAHNOB-HCI) |

To further analysis the brain regions in the HSLT, we visualize the heads within the self-attention. The average scores of the heads are illustrated in the Fig. 3. Each column can be seen as the contribution of the corresponding brain region or [CLS] in the classification. In the DEAP database, we find that the pre-frontal lobe and frontal lobe (PF and F in the Fig. 3(a) and Fig. 3(b)) have achieved more contributions than other regions. And similar finding could be observed in the valence recognition in the MAHNOB-HCI database (PF and F in the Fig. 3(c)). Our findings basically coincide with the observation of in the neuroscience, pre-frontal lobe and frontal lobe are closely correlated to the emotions [12]. And the parietal lobe (P in the Fig. 3(a), LP and RP in the Fig. 3(c)) achieves more contribution in the arousal recognition. This is basically consistent with the finding in the neuroscience study [13], the parietal lobe is associated with arousal changes in humans. Similarly, the findings of temporal lobe (LT in the Fig. 3(b) and Fig. 3(d)) in the valence recognition are also basically consistent with the findings of Koelstra *et al.* [26]. Besides, we notice a significant contribution from class token in the MAHNOB-HCI database ([CLS] in the Fig. 3(b) and Fig. 3(d)). It could verify the importance of the class token. Overall, these results indicate that the self-attention within HSLT could learn the discriminate features from the critical brain regions.

Furthermore, the Fig. 3. have shown the advantages of HSLT, and it also shown the difficulty in the EEG-based emotion recognition. We find the inconsistent emotional pattern between different subjects and databases. It is might associated with sensory information which is generated by different stimulus materials (Music videos in DEAP and movie clips in MAHNOB-HCI) and individual-specific differences [39].

### *C. Comparison With Related Researches*

We compare the performance of HSLT with recently researches using subject-independent scheme in the DEAP and MAHNOB-HCI database. The performances of different methods are listed in the Table XV. Except for the Ref. [41], all the studies adopt leave-one-subject-out cross-validation paradigm. The accuracy of HSLT is higher than the results in [27], [40], [41], [43] and [44]. And our method outperforms the Ref. [39] in the DEAP database and valence recognition in the MAHNOB-HCI. Compared with Ref. [42], our method is higher than it in the arousal recognition of DEAP database, but the performances of our method are less in the other tasks. Overall, the proposed HSLT has achieved the outstanding performance among these methods.

Compared with these works, the advantages of HSLT can be concluded as follows. Firstly, the self-attention mechanism emphasizes the contributive brain regions and enhances spatial dependencies capturing. Next, the hierarchical learning strategy could effectively integrate the spatial information within the brain regions and capture the dependencies among the brain regions. Finally, the PE and [CLS] retain the positional information and representative information which are essential for the sequence learning.

Meanwhile, the limitation of HSLT is also valuable to discuss. It is difficult to acquire a large amount of emotional EEG data [45]. However, transformer encoder needs the sufficient data for adequately revealing the strong ability of sequence learning. Although we have adopted sliding window to divide a trail into several segments for obtaining more samples, the data size restricts the performance of HSLT to some extent. For example, the performance would be slightly decrease when we design a deeper network than the proposed HSLT (in the Section V *HSLT with different configurations*).

## VII. Conclusion

In this work, we propose a novel HSLT model to robustly learn the EEG spatial information and apply it to achieve arousal and valence recognition. The PSD features are adopted to evaluate the HSLT performance, and the HSLT has achieved outstanding performance in subject-independent experiment with the accuracy of arousal and valence 65.75%/66.20% and 66.51%/66.63% in the DEAP and MAHNOB-HCI database respectively. We conduct extensive experiments to verify the effectiveness of HSLT. What's more, the visualization of the self-attention demonstrates that the self-attention within HSLT could emphasize discriminative information from contributive brain regions. We also find that positional embedding and class token also make contributions to boosting the performance. In addition, we also analysis the limitation of HSLT. In the future work, the data augmentation algorithms for EEG signals are necessary to investigate for further improving the emotion recognition performance.

## References

[1] P. Leelaarporn *et al.*, "Sensor-driven achieving of smart living: A review," *IEEE Sensors J.*, vol. 21, no. 9, pp. 10369–10391, May 2021.

[2] R. W. Picard, *Affective Computing*, vol. 604. Cambridge, MA, USA: MIT Press, 1997, pp. 247–248.

[3] F. Demir, N. Sobahi, S. Siuly, and A. Sengur, "Exploring deep learning features for automatic classification of human emotion using EEG rhythms," *IEEE Sensors J.*, vol. 21, no. 13, pp. 14923–14930, Jul. 2021.

[4] S. K. Khare and V. Bajaj, "An evolutionary optimized variational mode decomposition for emotion recognition," *IEEE Sensors J.*, vol. 21, no. 2, pp. 2035–2042, Jan. 2021.

[5] P. Lakhan *et al.*, "Consumer grade brain sensing for emotion recognition," *IEEE Sensors J.*, vol. 19, no. 21, pp. 9896–9907, Nov. 2019.

[6] L. Kulke, D. Feyerabend, and A. Schacht, "A comparison of the affectiva iMotions facial expression analysis software with EMG for identifying facial expressions of emotion," *Frontiers Psychol.*, vol. 11, p. 329, Feb. 2020.

[7] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG pattern analysis for emotion detection," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 102–115, Jan. 2012.

[8] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, Feb. 2013.

[9] X. W. Wang, D. Nie, and B. L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, no. 10, pp. 94–106, Apr. 2014.

[10] E. Kroupi, J. M. Vesin, and T. Ebrahimi, "Subject-independent odor pleasantness classification using brain and peripheral signals," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 422–434, Oct./Dec. 2016.

[11] S. K. Khare and V. Bajaj, "Time–frequency representation and convolutional neural network-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021.

[12] A. Etkin, T. Egner, and R. Kalisch, "Emotional processing in anterior cingulate and medial prefrontal cortex," *Trends Cogn. Sci.*, vol. 15, no. 2, pp. 85–93, Feb. 2011.

[13] S. Anders, M. Lotze, M. Erb, W. Grodd, and N. Birbaumer, "Brain activity underlying emotional valence and arousal: A response-related fMRI study," *Hum. Brain Mapping*, vol. 23, no. 4, pp. 200–209, 2004.

[14] S. Sangnark *et al.*, "Revealing preference in popular music through familiarity and brain response," *IEEE Sensors J.*, vol. 21, no. 13, pp. 14931–14940, Jul. 2021.

[15] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Jan. 2021, pp. 1–21.

[16] R. Khosrowabadi, C. Quek, K. K. Ang, and A. Wahab, "ERNN: A biologically inspired feedforward neural network to discriminate emotion from EEG signal," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 609–620, Mar. 2014.

[17] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, May 2016, pp. 1–15.

[18] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Mar. 2019.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[20] W. Tao *et al.*, "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Trans. Affect. Comput.*, early access, Sep. 22, 2020, doi: 10.1109/TAFFC.2020.3025777.

[21] J. X. Chen, D. M. Jiang, and Y. N. Zhang, "A hierarchical bidirectional GRU model with attention for eeg-based emotion classification," *IEEE Access*, vol. 7, pp. 118530–118540, 2019.

[22] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Ann. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, pp. 5998–6008.

[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (HAACL-HLT)*, Jun. 2019, pp. 4171–4186.

[24] J. Sun, J. Xie, and H. Zhou, "EEG classification with transformer-based models," in *Proc. IEEE 3rd Global Conf. Life Sci. Technol. (LifeTech)*, Mar. 2021, pp. 92–93.

[25] J. Pedoeem, S. Abittan, G. B. Yosef, and S. Keene, "TABS: Transformer based seizure detection," in *Proc. IEEE Signal Process. Med. Biol. Symp. (SPMB)*, Dec. 2020, pp. 1–6.

[26] S. Koelstra *et al.*, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[27] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[28] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[29] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barret, "The brain basis of emotion: A meta-analytic review," *Behav. Brain Sci.*, vol. 35, no. 3, pp. 121–143, Jun. 2012.

[30] K. Vytal and S. Hamann, "Neuroimaging support for discrete neural correlates of basic emotions: A voxel-based meta-analysis," *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2864–2885, Dec. 2010.

[31] G. C. Ribas, "The cerebral sulci and gyri," *Neurosurg. Focus*, vol. 28, no. 2, p. E2, Feb. 2010.

[32] W. Song *et al.*, "AutoInt: Automatic feature interaction learning via self-attentive neural networks," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Nov. 2019, pp. 1161–1170.

[33] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[34] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–15.

[35] S. Alhagry, A. A. Fahmy, 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. 355–358, 2017.

[36] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using deep belief networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

[37] Z. Gao, T. Yuan, X. Zhou, C. Ma, K. Ma, and P. Hui, "A deep learning method for improving the classification accuracy of SSMVEP-based BCI," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 12, pp. 3447–3451, Dec. 2020.

[38] 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. Cogn. Develop. Syst.*, vol. 13, no. 4, pp. 945–954, Dec. 2021.

[39] S. Rayatdoost and M. Soleymani, "Cross-corpus EEG-based emotion recognition," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2018, pp. 1–6.

[40] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, and B. Hu, "Exploring EEG features in cross-subject emotion recognition," *Frontiers Neurosci.*, vol. 12, p. 162, May 2018.

[41] P. Pandey and K. R. Seeja, "Subject independent emotion recognition from EEG using VMD and deep learning," *J. King Saud Univ. Comput. Inf. Sci.*, Nov. 2019.

[42] Z. Yin, L. Liu, J. Chen, B. Zhao, and Y. Wang, "Locally robust EEG feature selection for individual-independent emotion recognition," *Expert Syst. Appl.*, vol. 162, Dec. 2020, Art. no. 113768.

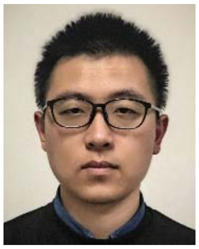
[43] W. Zhang, Z. Yin, Z. Sun, Y. Tian, and Y. Wang, "Selecting transferrable neurophysiological features for inter-individual emotion recognition via a shared-subspace feature elimination approach," *Comput. Biol. Med.*, vol. 123, Aug. 2020, Art. no. 103875.

[44] X. Huang *et al.*, "Multi-modal emotion analysis from facial expressions and electroencephalogram," *Comput. Vis. Image Understand.*, vol. 147, pp. 114–124, Jun. 2016.

[45] W. Li, W. Huan, B. Hou, Y. Tian, Z. Zhang, and A. Song, "Can emotion be transferred?—A review on transfer learning for EEG-based emotion recognition," *IEEE Trans. Cogn. Develop. Syst.*, early access, Jul. 21, 2021, doi: 10.1109/TCDS.2021.3098842.

**Zhe Wang** received the M.S. degree from the School of Electrical and Electronic Engineering, Tianjin University of Technology, China. He is currently pursuing the Ph.D. degree with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His current research interests include affective computing, signal processing, and brain–computer interface.

**Yongxiong Wang** received the B.S. degree in engineering mechanics from Harbin Engineering University, China, and the M.S. and Ph.D. degrees in control science and engineering from Shanghai Jiao Tong University, China. He is currently a Professor with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His research interests include computer vision, affective computing, and intelligent robot.



**Chuanfei Hu** received the M.S. degree from the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology. He is currently pursuing the Ph.D. degree with the School of Automation, Southeast University, China. His current research interests include deep learning and affective computing.

**Zhong Yin** received the Ph.D. degree in control science and engineering from the East China University of Science and Technology, China. He is currently an Associate Professor with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His research interests include intelligent human–machine systems, biomedical signal processing, and pattern recognition.

**Yu Song** (Member, IEEE) received the Ph.D. degree in intelligent mechanical systems engineering from Kagawa University, Japan. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Tianjin University of Technology, China. His current research interests include human–robot interaction, affective computing, brain–computer interface, and artificial intelligence.