TYPE Original Research
PUBLISHED 26 February 2025
DOI 10.3389/fnhum.2025.1517273

Check for updates

OPEN ACCESS

EDITED BY
Jiahui Pan,
South China Normal University, China

REVIEWED BY
Zhipeng He,
Sun Yat-sen University, China
Francesca Gasparini,
University of Milano-Bicocca, Italy

*CORRESPONDENCE
Yang Shen
✉ shenyang@bnu.edu.cn

RECEIVED 29 October 2024
ACCEPTED 10 February 2025
PUBLISHED 26 February 2025

CITATION
Li M, Yu P and Shen Y (2025) A spatial and temporal transformer-based EEG emotion recognition in VR environment.
Front. Hum. Neurosci. 19:1517273.
doi: 10.3389/fnhum.2025.1517273

COPYRIGHT
© 2025 Li, Yu and Shen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A spatial and temporal transformer-based EEG emotion recognition in VR environment

Ming Li[1], Peng Yu[1] and Yang Shen[2]*

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China,
[2]Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

With the rapid development of deep learning, Electroencephalograph(EEG) emotion recognition has played a significant role in affective brain-computer interfaces. Many advanced emotion recognition models have achieved excellent results. However, current research is mostly conducted in laboratory settings for emotion induction, which lacks sufficient ecological validity and differs significantly from real-world scenarios. Moreover, emotion recognition models are typically trained and tested on datasets collected in laboratory environments, with little validation of their effectiveness in real-world situations. VR, providing a highly immersive and realistic experience, is an ideal tool for emotional research. In this paper, we collect EEG data from participants while they watched VR videos. We propose a purely Transformer-based method, EmoSTT. We use two separate Transformer modules to comprehensively model the temporal and spatial information of EEG signals. We validate the effectiveness of EmoSTT on a passive paradigm collected in a laboratory environment and an active paradigm emotion dataset collected in a VR environment. Compared with state-of-the-art methods, our method achieves robust emotion classification performance and can be well transferred between different emotion elicitation paradigms.

KEYWORDS

electroencephalograph, virtual reality, transformer, emotion recognition, brain-computer interface

1 Introduction

Affective Brain-Computer Interfaces (aBCIs) is a system that inducts, recognize, and regulate human emotions, which involves computer science, psychology, cognitive science, and more, aiming to enhance computers' ability to understand and respond to human emotional states during human-computer interaction (Tao and Tan, 2005). The Inputs of aBCIs typically include functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS), and electroencephalography (EEG). Among these, EEG non-invasively captures electrical activity on the scalp, offering high temporal resolution and relatively easy signal acquisition (Alarcao and Fonseca, 2017). Consequently, it has been widely applied in medical rehabilitation, education, and other fields (Zheng et al., 2023; Li et al., 2020). In the field of medical rehabilitation, aBCI can objectively and accurately assess emotional states, providing a valuable supplement to traditional diagnostic methods that are highly subjective, such as behavioral observations and questionnaires. In transportation, EEG signals can monitor drivers' negative emotional states like anxiety and anger in real time, providing timely warnings and adjustments to reduce traffic accidents (Zepf et al., 2020).

In these applications, emotion recognition is the most critical component (Hu et al., 2019). However, due to EEG signals' sensitivity to noise, numerous artifacts are introduced, posing significant challenges to emotion recognition. Traditional machine learning models extract features related to emotional states from EEG signals, such as power spectral density (PSD) (Solomon Jr, 1991) and differential entropy (DE) (Duan et al., 2013), and then feed them into classifiers like SVM (Wang et al., 2011), KNN (Mehmood and Lee, 2015), and MLP (Li et al., 2022a), achieving decoding of EEG. With the advancement of deep learning, researchers have developed various models to decode emotions from EEG, including supervised emotion recognition models based on convolutional neural networks (CNN) (Hwang et al., 2020), recurrent neural networks (RNN) (Alhagry et al., 2017), graph convolutional neural networks (GCN) (Song et al., 2018), and Transformers (Sun et al., 2022), which have made significant progress. CNNs perform well in classification tasks, particularly in fields like image (Bhatt et al., 2021), video (Xu et al., 2015), and speech processing (Hema and Marquez, 2023). RNNs excel at handling sequential data but face limitations in parallel training and global information capture (Ma et al., 2023). In contrast, Transformer models utilize self-attention mechanisms to effectively capture crucial long-term dependencies in time series (Chitty-Venkata et al., 2023).

Initially applied in natural language processing and computer vision with remarkable results, Transformers have recently begun to be employed in EEG encoding and decoding tasks (Abibullaev et al., 2023). By capturing long-term temporal relationships in EEG sequences, they extract robust feature representations. However, most studies focus solely on modeling either the temporal or spatial dimensions of EEG, allowing for the learning of relationships between different channels or time frames (Peng et al., 2023). When using a temporal-spatial Transformer, each channel of each frame is treated as a token. This approach leads to a significant increase in the number of tokens when processing long EEG sequences or multi-channel data, resulting in substantial computational demands.

Moreover, in the field of affective computing, emotional induction paradigms can be divided into laboratory settings (passive induction paradigms) and natural settings (active induction paradigms) (Meuleman and Rudrauf, 2021). However, current research is almost exclusively conducted in laboratory settings, where the passively induced emotional changes differ from the actively generated emotional changes in real-world scenarios (Miranda-Correa et al., 2018; Katsigiannis and Ramzan, 2017). Moreover, most emotion recognition studies have focused on training and testing models on a single dataset within a laboratory environment, with few studies validating the effectiveness of emotion recognition models in natural settings (Marín-Morales et al., 2020). Virtual reality (VR) can provide a highly immersive and realistic virtual environment, allowing for the assessment of emotional experiences in a more realistic scenario, making it an ideal paradigm for affective research (Marín-Morales et al., 2020).

In this paper, we conducted an emotional induction experiment in a VR environment and collected corresponding EEG data. We propose an emotion recognition method based on spatial and temporal Transformers (EmoSTT). EmoSTT employs two separate Transformers to model the temporal and spatial dimensions of EEG data, without being affected by the excessive number of tokens caused by long time sequences or multi-channel EEG data. The temporal and spatial Transformer blocks can learn the correlations between EEG time series and across channels, extracting hidden feature representations with spatial-temporal dependencies. Finally, the feature representations are fed into a simple fully connected layer to decode emotional states. Lastly, we validated the model's effectiveness on datasets of different emotional induction paradigms.

2 Related work

2.1 EEG-based emotion recognition

Traditional machine learning-based emotion recognition typically involves preprocessing, feature extraction, feature smoothing, training classifiers, and testing (Jenke et al., 2014). Signal features can be categorized into time-domain, frequency-domain, and spatial-domain features. Among these, frequency-domain features are most relevant to emotions, and DE features have been proven to offer the best emotion recognition performance (Jenke et al., 2014). It commonly use the Short Time Fourier Transform (STFT) to convert sequential EEG signals into the frequency domain, thereby extracting features from five frequency bands: δ band (1–3Hz), θ band (4–7Hz), α band (8–13Hz), β band (14–30Hz), and γ band (31–50Hz) (Mohammadi et al., 2017). Additionally, deep learning models are dedicated to designing neural networks that extract generalizable features. Li et al. (2016) further proposed a hybrid deep learning architecture (Convolutional and Recurrent Neural Network, C-RNN) for emotion recognition. The model extracts task-relevant features, explores channel correlations, and integrates contextual information across these frames, achieving excellent results on the DEAP dataset (Koelstra et al., 2011). Li et al. (2019) introduced a Spatial-Temporal Neural Network with Regional to Global (R2G-STNN) based on Bidirectional Long Short-Term Memory (BiLSTM), which conducts hierarchical feature learning from regional to global through spatial and temporal neural network models to extract discriminative spatiotemporal EEG features. Zhong et al. (2020) proposed a Regularized Graph Neural Network (RGNN) that considers the biological topology between different brain regions to capture local and global relationships between different EEG channels. Additionally, two regularizers were proposed, namely Node Domain Adversarial Training (NodeDAT) and Emotion-Aware Distribution Learning (EmotionDL), to better handle individual differences and noisy label issues. Song et al. (2022) proposed a novel compact convolutional Transformer network called EEG-Conformer for improving emotion recognition performance of EEG signals. EEG-Conformer combines convolutional modules to capture local temporal and spatial features, as well as self-attention modules to extract global correlations.

2.2 Active and passive emotion elicitation paradigms

However, the aforementioned methods are all trained and tested on data collected in traditional laboratory settings. These represent passive emotional changes, which differ from the active emotional changes that individuals generate in real-world scenarios, potentially leading to differences in EEG signals between these two paradigms (Somarathna et al., 2022). The ecological validity in emotional induction is crucial for affective research (Mohammadi and Vuilleumier, 2020). The passive emotional induction paradigm in traditional laboratory settings has weak induction effects (Soleymani et al., 2011). In contrast, virtual reality (VR) can simulate controlled environments with high immersion, presence, and interactivity, evoking emotions more naturally and authentically (Cao et al., 2021). Previous studies have validated the effectiveness of VR emotional induction paradigms (Li et al., 2022b). However, due to the lack of active paradigm emotional EEG datasets, research on emotion recognition models in active paradigms is very limited.

3 Methods

3.1 Pipeline

The overall framework of the model is depicted in Figure 1. Initially, the raw EEG signals undergo preprocessing, followed by segmenting all signals into 1-second epochs. Consistent with (Duan et al., 2013), for each 1-second segment, we employ the Short-Time Fourier Transform (STFT) for frequency domain feature extraction. We extract the DE features from the EEG data of all participants $\mathbf{X} \in \mathbb{R}^{N\times C\times F}$ as the pre-training dataset to input our model, where N represents the number of samples in the preprocessed EEG dataset, C refers to the number of EEG channels, and F is the dimensionality of the DE features. Here, we set F = 5, corresponding to the five frequency bands: δ band (1–3Hz), θ band (4–7Hz), α band (8–13Hz), β band (14–30Hz), and γ band (31–50Hz). To better capture the temporal dimension of the EEG data, we use an overlapping time window of length T to transform the original signal $\mathbf{X}$ into the shape $\mathbf{X}_{new} \in \mathbb{R}^{N_{new}\times T\times C\times F}$, which serves as the input for the final pre-training model, each sample is represented as $\mathbf{x}_s \in \mathbb{R}^{T\times C\times F}$ Here, N_{new} represents the number of samples in the dataset. T denotes the number of time frames, which we set to 10, consistent with previous classical studies (Li et al., 2022c) to facilitate comparison. To mitigate the variability among features and enhance performance, we normalize training and testing data based on the mean and standard deviation of the training set. Subsequently, we utilize two separate Transformers to extract the spatial and temporal features of the EEG signals. Below is a detailed description of the method.

3.2 Spatial transformer

For the input to the spatial transformer $\mathbf{X}_{new} \in \mathbb{R}^{N_{new}\times T\times C\times F}$, where each sample is represented as $\mathbf{X}_{new} \in \mathbb{R}^{T\times C\times F}$. Then, we then project the frequency domain dimension K to a hidden dimension D using a linear projection matrix $P \in \mathbb{R}^{k\times D}$.

Since EEG signals record complex brain activities from multiple electrode channels, there is a strong correlation between different electrode channels. Therefore, using a spatial transformer can effectively encode the spatial information of EEG signals. For each given frame of data $\{\mathbf{x}_s^i \in \mathbb{R}^{C\times D} \mid$i = 1, 2, ..., T$\}$, where C is the number of channels, each channel is treated as a patch, and a learnable spatial position encoding $\mathbf{E}_{SPos} \in \mathbb{R}^{C\times D}$is added to $\mathbf{x}_s^i$. This preserves the position information of each channel, which is crucial in transformers and is a standard and common practice. Here, we use learnable sine-cosine trigonometric functions as spatial position encoding. We utilize the attention mechanism to extract the functional connectivity relationships between different electrode channels by stacking multiple transformer blocks. The computation process of self-attention is as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In this context, Q, K, and V represent the query, key, and value, respectively. For each input vector, three linear transformations are applied to map it into the query vector, key vector, and value vector. For each query vector, the similarity to all key vectors is calculated using a dot product, resulting in attention scores. To prevent the dot product from becoming too large, it is divided by the square root of d_k; subsequently, the Softmax function is applied to normalize the results of the aforementioned dot product. Finally, after obtaining the Softmax matrix, it is multiplied by V to produce the final output. Here, we utilize the multi-head attention mechanism (MHA), which captures dependencies in different subspaces of the input sequence by computing multiple sets of queries, keys, and values in parallel. The outputs of these heads are then concatenated and passed through a linear transformation to yield the final output. As illustrated in Figure 2, in addition to the multi-head attention mechanism, each Transformer encoder includes multi-layer perceptrons (MLPs), with each component employing residual connections and layer normalization to enhance the model's training efficiency and performance.

3.3 Temporal transformer

Subsequently, the output from the spatial Transformer module is fed into the temporal encoder module. We add the same sine-cosine positional encoding to keep track of the position of each time frame. Like the spatial Transformer, we stack the same number of Transformer blocks. Unlike RNN, transformers are better at capturing long-term dependencies in EEG signals. The encoder also includes multi-head attention (MHA) and multi-layer perceptrons (MLPs). Finally, for the output of the temporal Transformer, we pass it through a simple MLP block with layer normalization and a linear layer to obtain the final classification output $\mathbf{y} \in \mathbb{R}^{C\times K}$. In this paper, we use cross-entropy loss to minimize the error between the predicted emotion categories and the true emotion labels:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_n^k \log\left(\hat{y}_n^k\right) \tag{2}$$

where N represents the number of batch sizes and K represents the number of categories. y_n^k is the true emotion label and $\hat{y}_n^k$ is the predicted emotion category.

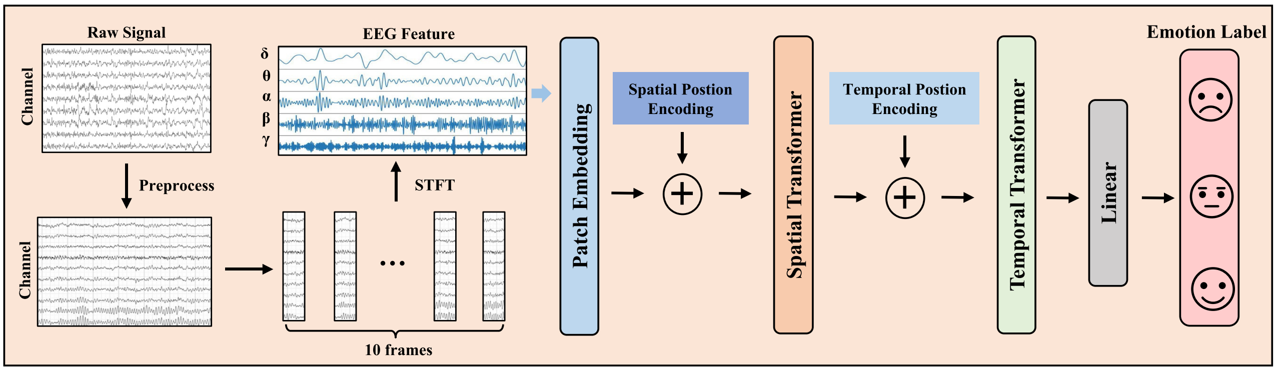

FIGURE 1
The overall model framework of EmoSTT.

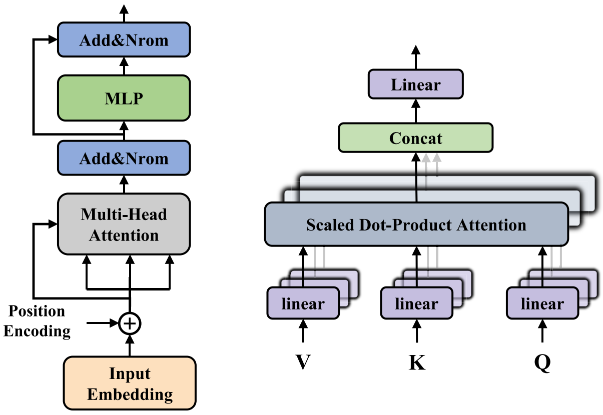

FIGURE 2
The left is the overall architecture of the transformer, and the right is the multi-head attention module.

4 Experiments

4.1 Dataset

This paper validates the proposed method on a widely used public dataset, SEED (Zheng and Lu, 2015), as well as on a self-collected VR emotion dataset (VR-Emotion). Both datasets consist of EEG signals collected from participants while they watch video stimuli in a quiet laboratory or a VR environment. After viewing the video stimuli, participants provide self-assessments of their emotional responses, which serve as the labels for the EEG data. Below is a brief introduction to the datasets:

4.1.1 SEED dataset

The SEED dataset primarily comprises EEG signals corresponding to three types of emotions: positive, neutral, and negative. The data were collected using a 62-channel device from the ESI NeuroScan system, with a sampling rate of 1,000 Hz. To account for the impact of cultural differences on emotion recognition research, all movie clips selected for this dataset are in Mandarin. The three emotion experiments involved 15 participants (7 males and 8 females), all Chinese students from Shanghai Jiao Tong University. To verify the stability of emotions over different periods, each participant experimented three times, with a one-week interval between each session, resulting in a total of 45 experiments. Participants were required to watch 15 emotional stimulus videos in each experiment, corresponding to the three emotion categories. Each video ranged from 185 to 238 seconds in length. Each second of EEG data corresponding to the video is considered as one sample, so there are 3394 samples in each session.

4.1.2 VR-emotion dataset

In previous experiments, we have verified the effectiveness of the VR emotion induction paradigm (Li et al., 2022b). We recorded EEG signals from participants using the ANT Neuro EEG acquisition system while they watched VR video stimuli. The following principles guided the selection of VR video stimuli: (1) Each video should not be too long to avoid causing mental fatigue in participants; (2) The content must be clear and understandable without the need for translation; (3) The movie clips must evoke a single target emotional state. Therefore, we selected 4 videos from the Stanford public VR video dataset [2 for low valence/low arousal (LVLA) and 2 for high valence/low arousal (HVLA)] (Li et al., 2017). Due to the lack of high arousal/low valence (HALV) videos, we chose 15 of YouTube's most viewed horror videos. We then invited 16 students majoring in psychology from Beijing Normal University to rate the videos' emotional arousal and valence dimensions. For each video x, we calculated the normalized arousal and valence scores by dividing the average score by the standard deviation (μ_x/σ_x). Ultimately, we selected two horror videos with extreme angles in the VA plane quadrant: "Real Run" and "The Conjuring 2." For the high arousal/high valence (HAHV) videos, during the pre-experiment, all participants reported severe motion sickness from these two videos, which could affect EEG analysis (Jeong et al., 2019). To eliminate the impact of motion sickness in the experiment, we chose the VR game "The Blu" from the Steam platform. This game offers a more passive and intuitive experience, akin to watching a video, and has been proven to evoke high arousal/positive valence emotions while providing an immersive VR experience (Meuleman and Rudrauf, 2021). The game includes three segments: "The Reef Migration," "The Whale Encounter," and "The Hammerhead Cove." We invited the same psychology students to evaluate these three segments and ultimately selected "The Whale Encounter" and "The Reef Migration" as the final emotional stimulus materials.

The dataset comprises EEG data from 28 participants (16 males and 12 females) collected while viewing 8 VR videos, using a 32-channel EEG system with a sampling rate of 512 Hz. Each video is approximately 3 min and 30 seconds long. EEG data corresponding to every second of each video is considered a sample, amounting to 1,483 samples per participant.

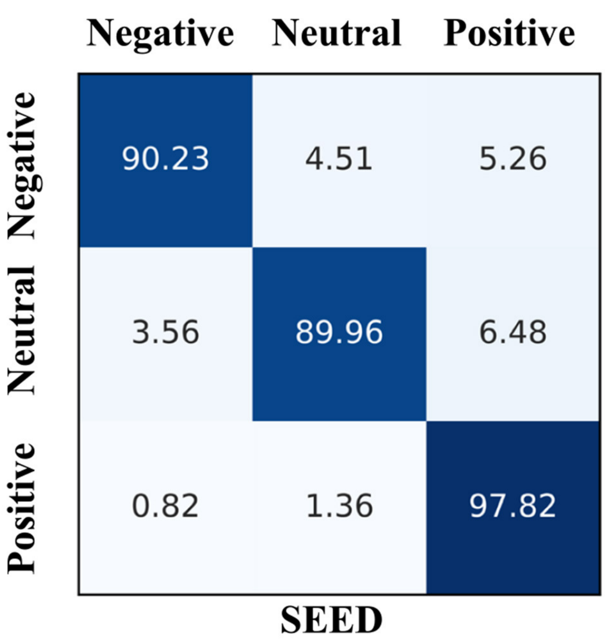

FIGURE 3
The confusion matrix of SEED dataset classification results.

4.2 EEG preprocessing and feature extraction

The authors provided the original preprocessed DE features for the SEED dataset, which we directly use as input in this chapter. As for the VR-Emotion dataset, to obtain clean and high-quality EEG signals, we require participants to avoid excessive head movements while recording EEG signals in VR to ensure signal stability and reliability. To prevent the HMD from exerting pressure on the front-central electrodes, a lateral elastic band is used to fix the HMD while the upper elastic band is loose. Additionally, to avoid the potential impact of pressure from repeatedly wearing the VR headset on the quality of the EEG signals, the subjective questionnaire is presented directly on the VR screen, enabling participants to complete it without removing the headset. For the raw EEG data, we employ EEGLAB for EEG signal processing (Delorme and Makeig, 2004). EEGLAB is an open-source Matlab toolbox that provides algorithms for EEG preprocessing and feature extraction. The specific steps are as follows: (1) First, the original EEG data has a sampling frequency of 512 Hz, which is sufficient to filter out interference from the monitor (50–60 Hz) and the VR headset (90 Hz). The signals are then downsampled to 128 Hz and re-referenced using bilateral mastoid electrodes (M1 and M2). (2) Since EEG signals are low-frequency and electromyographic artifacts are high-frequency, low-pass filtering removes the EMG artifacts significantly. Furthermore, a bandpass filter (4–47 Hz) is applied to the signal using the FIR filter in EEGLAB to filter out eye movement artifacts better. (3) Visual inspection is performed to remove abnormal signals with amplitudes exceeding $\pm 100\ \mu$V, as signals beyond this threshold are considered non-EEG signals. (4) Independent Component Analysis (ICA) (Chaumon et al., 2015) is then used to decompose the original signal into 32 independent components (ICs). ICA is a method based on Blind Source Separation (BSS). Using the SASICA plugin in EEGLAB and visual inspection, we identify which components are related to emotion and which are artifacts or other neural activity components (such as eye blinks, muscle activity, or head movements). (5) Finally, for each participant, an average of 9.37 components are removed, yielding clean EEG signals.

Then, we extract DE features based on the preprocessed data in the same way as the previous studies (Duan et al., 2013). DE has been proven to be one of the most effective features for emotion recognition (Qiu et al., 2024; He et al., 2022). DE features are an extended form of Shannon's information entropy $-\sum_x p(x)\log(p(x))\,dx$ for continuous variables, and they are calculated as follows:

$$DE = -\int_a^b p(x)\log(p(x))\,dx \tag{3}$$

In this context, $p(x)$ represents the probability density function of continuous information, and $|a, b|$ indicates the interval over which the information is taken. For a segment of EEG signals that approximately follow a Gaussian distribution $N(\mu, \sigma_i^2)$ for a specific length, its differential entropy equals the logarithm of its energy spectrum in a particular frequency band.

$$DE = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}}\right) dx = \frac{1}{2}\log(2\pi e\sigma_i^2) \tag{4}$$

4.3 Evaluation

To thoroughly evaluate the effectiveness of the emotion recognition pre-training framework, this chapter conducts subject-specific emotion recognition experiments on the SEED and VR-Emotion datasets, that is, training an emotion classification model separately for each individual's EEG data from each experimental session. Consistent with previous research methods (Zheng et al., 2018; Liu et al., 2021), the training, fine-tuning, and testing data come from a single session on the same subject. In the SEED dataset, the first 9 trials of each session are used as training data (with 3 trials for each emotion category to maintain data balance),

TABLE 1 The ACC/F1 of subject-dependent experiments on the SEED dataset.

Method	SVM	RF	DBN	GRSLR	GCNN
ACC/F1	83.99/78.93	78.46/76.58	86.08/79.89	87.39/80.74	87.40/81.33
Method	DGCNN	DANN	BiDANN	EmoSTT	
ACC/F1	90.40/84.85	91.36/85.26	92.38/86.28	**92.67/86.78**	

Bold indicates that our method achieved the best results.

and the last 6 trials are used as testing data. Consequently, there are a total of 45 emotion recognition models, and the final classification accuracy is defined as the average accuracy obtained from these 45 models. For the VR-Emotion dataset, the first 6 trials of each subject are used as the training set (3 for high arousal/low arousal and 3 for low valence/high valence, to maintain data balance), and the last 2 trials are used as the test set. Thus, there are a total of 28 emotion recognition models, and the final classification accuracy is defined as the average accuracy obtained from these 28 models.

TABLE 2 Comparison of accuracy with standard deviation between DGCNN and EmoSTT.

Subject	DGCNN	EmoSTT
1	89.39 ± 0.93	96.72 ± 6.8.67
2	80.00 ± 1.95	88.92 ± 8.21
3	83.85 ± 1.34	96.47 ± 6.54
4	94.01 ± 1.87	92.94 ± 7.24
5	85.12 ± 1.09	88.07 ± 5.64
6	91.45 ± 0.78	90.95 ± 7.51
7	91.45 ± 1.41	96.10 ± 5.89
8	87.77 ± 1.74	91.72 ± 7.67
9	94.37 ± 1.66	92.35 ± 6.53
10	82.99 ± 1.50	79.06 ± 7.98
11	92.10 ± 0.67	95.58 ± 6.83
12	90.04 ± 1.99	92.68 ± 7.09
13	90.66 ± 1.04	93.63 ± 6.46
14	92.60 ± 1.25	95.29 ± 8.29
15	97.79 ± 0.46	99.49 ± 7.85
Aver	90.40/08.49	92.67/7.05

TABLE 3 Comparison of models on VR-emotion dataset for arousal and valence.

Model	VR-emotion (arousal)	VR-emotion (valence)
SVM	64.58/60.34	68.36/64.21
RF	66.48/61.38	69.42/65.46
DGCNN	72.48/67.18	71.59/67.73
EmoSTT	75.67/70.83	76.47/71.29

The result is expressed as ACC/F1.

4.4 Implementation

This paper employs the PyTorch framework to train models on an NVIDIA 4090 GPU. The model utilizes four Transformer modules, with the number of attention heads set to 8 and the hidden dimension D set to 16. We adopt an exponential learning rate decay scheme with an initial learning rate of 2×10^{-4} and a decay factor of 0.98 per epoch. Weight decay and batch size are set to 0.1 and 128, respectively. The model is optimized using the AdamW optimizer. For both temporal and spatial encoder, we use $L = 6$ transformer blocks. The embedding dimension D is set to 32, and the multi-head number H is set to 6.

4.5 Results

4.5.1 SEED dataset

We validate the model's classification results on the SEED dataset using topic-related experiments. As shown in the Table 1, we compared our model with several other supervised models. The results indicate that our model achieved an accuracy of 92.67% and an F1 score of 86.78% for positive, negative, and neutral emotions, outperforming other supervised methods. This demonstrates that the model can extract robust and highly discriminative features.

Figure 3 presents the confusion matrix of EmoSTT's results on the SEED dataset, where each row represents the true class and each column represents the predicted class. The results indicate that positive emotions are the most easily identified, achieving an accuracy rate of 97.82%. The accuracy rates for neutral and negative emotions are 89.96% and 90.23%, respectively. Additionally, we observe that neutral and positive emotions are more likely to be confused, a mix-up that may stem from the difficulty participants face in distinguishing between these two emotions during the emotion induction process.

In addition, we also list the average accuracy of three experiments for each subject and compare them with DGCNN (Song et al., 2018), as shown in the Table 2. It can be seen that the model achieves a relatively high average accuracy on the 15 subjects of the SEED dataset, and the average accuracy of our proposed EmoSTT is higher.

4.5.2 VR-emotion dataset

The results of the VR-Emotion dataset are shown in Table 3. EmoSTT achieved accuracy of 75.67% and 76.47% in the arousal and valence dimensions, respectively. The corresponding F1 scores are 70.83% and 71.29%, respectively. It can be seen that the model can also maintain robust performance under the active emotion induction paradigm. For emotional arousal, it outperformed SVM,

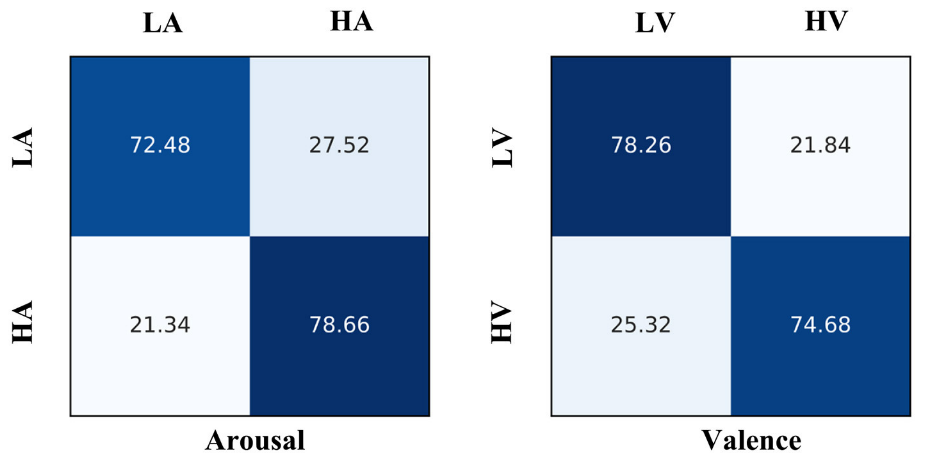

FIGURE 4
The confusion matrix of arousal and valence classification results of VR-emotion dataset.

RF, and DGCNN by 11.09%, 9.19%, and 3.21%, respectively. For emotional valence, it surpassed SVM, RF, and DGCNN by 8.11%, 7.05%, and 4.88%, respectively.

For the VR-Emotion dataset, Figure 4 shows the confusion matrix of the EmoSTT classification results. It can be seen that high-arousal emotions are relatively easy to identify, reaching an accuracy of 78.66%, which is 6.18% higher than the recognition of low-arousal emotions. We speculate that this is because VR videos are more advantageous in inducing high-arousal emotions. In terms of valence, the recognition accuracy of low valence is 78.26%, which is 3.58% higher than the 74.68% of high valence.

5 Conclusion and future work

In this paper, we propose EmoSTT, an emotion recognition model based on a pure Transformer model that can extract temporal and spatial dependency features of EEG signals. Taking advantage of the high ecological validity of the VR emotion induction paradigm, we collect an emotional EEG dataset of subjects watching VR videos. The performance of EmoSTT is verified on datasets of two different emotion induction paradigms. The model achieves an accuracy of 92.67% on the SEED passive induction dataset, and 75.67% and 76.47% arousal and valence classification accuracies on the VR-Emotion dataset, respectively. The results show that the model can well transfer the emotion recognition model in the laboratory environment to natural environments such as VR. In the future, we will validate our approach in a broader range of scenarios. This includes using more natural VR interactive environments, ensuring signal stability, and verifying the effectiveness of emotion induction as well as the robustness of the model.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: this dataset is an emotional EEG dataset. Requests to access these datasets should be directed to minglee@buaa.edu.cn.

Ethics statement

The studies involving humans were approved by the Biological and Medical Ethics Committee of Beihang University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

ML: Writing – original draft, Writing – review & editing. PY: Data curation, Formal analysis, Methodology, Software, Writing – review & editing. YS: Data curation, Formal analysis, Funding acquisition, Investigation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research is supported by the National Key R&D Program of China (No. 2022ZD0115902) and the National Natural Science Foundation of China (No.62377004, No. 62307003).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abibullaev, B., Keutayeva, A., and Zollanvari, A. (2023). Deep learning in EEG-based bcis: a comprehensive review of transformer models, advantages, challenges, and applications. *IEEE Access* 11, 127271–127301. doi: 10.1109/ACCESS.2023.3329678

Alarcao, S. M., and Fonseca, M. J. (2017). Emotions recognition using EEG signals: a survey. *IEEE Trans. Affect. Comput*. 10, 374–393. doi: 10.1109/TAFFC.2017.2714671

Alhagry, S., Fahmy, A. A., and El-Khoribi, R. A. (2017). Emotion recognition based on EEG using LSTM recurrent neural network. *Int. J. Adv. Comput. Sci. Applic*. 8:81046. doi: 10.14569/IJACSA.2017.081046

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., et al. (2021). Cnn variants for computer vision: History, architecture, application, challenges and future scope. *Electronics* 10:2470. doi: 10.3390/electronics10202470

Cao, R., Zou-Williams, L., Cunningham, A., Walsh, J., Kohler, M., and Thornas, B. H. (2021). "Comparing the neuro-physiological effects of cinematic virtual reality with 2D monitors," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (IEEE), 729–738. doi: 10.1109/VR50410.2021.00100

Chaumon, M., Bishop, D. V., and Busch, N. A. (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *J. Neurosci. Methods* 250, 47–63. doi: 10.1016/j.jneumeth.2015.02.025

Chitty-Venkata, K. T., Mittal, S., Emani, M., Vishwanath, V., and Somani, A. K. (2023). A survey of techniques for optimizing transformer inference. *J. Systems Archit*. 144:102990. doi: 10.1016/j.sysarc.2023.102990

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Duan, R.-N., Zhu, J.-Y., and Lu, B.-L. (2013). "Differential entropy feature for EEG-based emotion classification," in *2013 6th international IEEE/EMBS conference on neural engineering (NER)* (IEEE), 81–84. doi: 10.1109/NER.2013.6695876

He, Z., Zhong, Y., and Pan, J. (2022). An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition. *Comput. Biol. Med*. 141:105048. doi: 10.1016/j.compbiomed.2021.105048

Hema, C., and Marquez, F. P. G. (2023). Emotional speech recognition using CNN and deep learning techniques. *Appl. Acoust*. 211:109492. doi: 10.1016/j.apacoust.2023.109492

Hu, X., Chen, J., Wang, F., and Zhang, D. (2019). Ten challenges for EEG-based affective computing. *Brain Sci. Adv*. 5, 1–20. doi: 10.1177/2096595819896200

Hwang, S., Hong, K., Son, G., and Byun, H. (2020). Learning cnn features from de features for EEG-based emotion recognition. *Pattern Anal. Applic*. 23, 1323–1335. doi: 10.1007/s10044-019-00860-w

Jenke, R., Peer, A., and Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput*. 5, 327–339. doi: 10.1109/TAFFC.2014.2339834

Jeong, D., Yoo, S., and Yun, J. (2019). "Cybersickness analysis with EEG using deep learning algorithms," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (IEEE), 827–835. doi: 10.1109/VR.2019.8798334

Katsigiannis, S., and Ramzan, N. (2017). Dreamer: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform*. 22, 98–107. doi: 10.1109/JBHI.2017.2688239

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2011). Deap: a database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput*. 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Li, B. J., Bailenson, J. N., Pines, A., Greenleaf, W. J., and Williams, L. M. (2017). A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Front. Psychol*. 8:2116. doi: 10.3389/fpsyg.2017.02116

Li, C., Zhang, Z., Zhang, X., Huang, G., Liu, Y., and Chen, X. (2022a). EEG-based emotion recognition via transformer neural architecture search. *IEEE Trans. Ind. Inform*. 19, 6016–6025. doi: 10.1109/TII.2022.3170422

Li, L., Gow, A., and Zhou, J. (2020). The role of positive emotions in education: a neuroscience perspective. *Mind, Brain, Educ*. 14, 220–234. doi: 10.1111/mbe.12244

Li, M., Pan, J., Gao, Y., Shen, Y., Luo, F., Dai, J., et al. (2022b). Neurophysiological and subjective analysis of VR emotion induction paradigm. *IEEE Trans. Vis. Comput. Graph*. 28, 3832–3842. doi: 10.1109/TVCG.2022.3203099

Li, R., Wang, Y., Zheng, W.-L., and Lu, B.-L. (2022c). "A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning," in *Proceedings of the 30th ACM International Conference on Multimedia*, 6–14. doi: 10.1145/3503161.3548243

Li, X., Song, D., Zhang, P., Yu, G., Hou, Y., and Hu, B. (2016). "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE), 352–359. doi: 10.1109/BIBM.2016.7822545

Li, Y., Zheng, W., Wang, L., Zong, Y., and Cui, Z. (2019). From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput*. 13, 568–578. doi: 10.1109/TAFFC.2019.2922912

Liu, W., Qiu, J.-L., Zheng, W.-L., and Lu, B.-L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Trans. Cogn. Dev. Syst*. 14, 715–729. doi: 10.1109/TCDS.2021.3071170

Ma, Z., Zhang, H., and Liu, J. (2023). Mm-RNN: a multimodal rnn for precipitation nowcasting. *IEEE Trans. Geosci. Rem. Sens*. 61, 1–14. doi: 10.1109/TGRS.2023.3264545

Marín-Morales, J., Llinares, C., Guixeres, J., and Alcañiz, M. (2020). Emotion recognition in immersive virtual reality: from statistics to affective computing. *Sensors* 20:5163. doi: 10.3390/s20185163

Mehmood, R. M., and Lee, H. J. (2015). "Emotion classification of EEG brain signal using svm and knn," in *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)* (IEEE), 1–5. doi: 10.1109/ICMEW.2015.7169786

Meuleman, B., and Rudrauf, D. (2021). Induction and profiling of strong multi-componential emotions in virtual reality. *IEEE Trans. Affect. Comput*. 12, 189–202. doi: 10.1109/TAFFC.2018.2864730

Miranda-Correa, J. A., Abadi, M. K., Sebe, N., and Patras, I. (2018). Amigos: a dataset for affect, personality and mood research on individuals and groups. *IEEE Trans. Affect. Comput*. 12, 479–493. doi: 10.1109/TAFFC.2018.2884461

Mohammadi, G., and Vuilleumier, P. (2020). A multi-componential approach to emotion recognition and the effect of personality. *IEEE Trans. Affect. Comput*. 13, 1127–1139. doi: 10.1109/TAFFC.2020.3028109

Mohammadi, Z., Frounchi, J., and Amiri, M. (2017). Wavelet-based emotion recognition system using EEG signal. *Neural Comput. Applic*. 28, 1985–1990. doi: 10.1007/s00521-015-2149-8

Peng, G., Zhao, K., Zhang, H., Xu, D., and Kong, X. (2023). Temporal relative transformer encoding cooperating with channel attention for EEG emotion analysis. *Comput. Biol. Med*. 154:106537. doi: 10.1016/j.compbiomed.2023.106537

Qiu, L., Zhong, L., Li, J., Feng, W., Zhou, C., and Pan, J. (2024). SFT-SGAT: a semi-supervised fine-tuning self-supervised graph attention network for emotion recognition and consciousness detection. *Neural Netw*. 180:106643. doi: 10.1016/j.neunet.2024.106643

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput*. 3, 42–55. doi: 10.1109/T-AFFC.2011.25

Solomon Jr, O. M. (1991). Psd computations using welch's method. *NASA STI/Recon. Techn. Rep*. 92:23584.

Somarathna, R., Bednarz, T., and Mohammadi, G. (2022). Virtual reality for emotion elicitation-a review. *IEEE Trans. Affect. Comput*. 14, 2626–2645. doi: 10.1109/TAFFC.2022.3181053

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput*. 11, 532–541. doi: 10.1109/TAFFC.2018.2817622

Song, Y., Zheng, Q., Liu, B., and Gao, X. (2022). EEG conformer: convolutional transformer for EEG decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng*. 31, 710–719. doi: 10.1109/TNSRE.2022.3230250

Sun, M., Cui, W., Yu, S., Han, H., Hu, B., and Li, Y. (2022). A dual-branch dynamic graph convolution based adaptive transformer feature fusion network for EEG emotion recognition. *IEEE Trans. Affect. Comput*. 13, 2218–2228. doi: 10.1109/TAFFC.2022.3199075

Tao, J., and Tan, T. (2005). "Affective computing: a review," in *International Conference on Affective Computing and Intelligent Interaction* (Springer), 981–995. doi: 10.1007/11573548_125

Wang, X.-W., Nie, D., and Lu, B.-L. (2011). "EEG-based emotion recognition using frequency domain features and support vector machines," in *Neural Information Processing: 18th International Conference, ICONIP 2011, Shanghai, China, November 13–17, 2011, Proceedings, Part I 18* (Springer), 734–743. doi: 10.1007/978-3-642-24955-6_87

Xu, Z., Yang, Y., and Hauptmann, A. G. (2015). "A discriminative cnn video representation for event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1798–1807. doi: 10.1109/CVPR.2015.7298789

Zepf, S., Hernandez, J., Schmitt, A., Minker, W., and Picard, R. W. (2020). Driver emotion recognition for intelligent vehicles: a survey. *ACM Comput. Surv*. 53, 1–30. doi: 10.1145/3388790

Zheng, W., Yan, L., and Wang, F.-Y. (2023). Two birds with one stone: Knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. *IEEE Trans. Affect. Comput*. 14, 2595–2613. doi: 10.1109/TAFFC.2023.3282704

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., and Cichocki, A. (2018). Emotionmeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybern*. 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zheng, W.-L., and Lu, B.-L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev*. 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zhong, P., Wang, D., and Miao, C. (2020). Eeg-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comput*. 13, 1290–1301. doi: 10.1109/TAFFC.2020.2994159