Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

VG-DOCoT: a novel DO-Conv and transformer framework via VAE-GAN technique for EEG emotion recognition*

Yanping ZHU‡, Lei HUANG, Jixin CHEN, Shenyun WANG, Fayu WAN, Jianan CHEN

School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China

E-mail: 001520@nuist.edu.cn; 20211249221@nuist.edu.cn; 202212490689@nuist.edu.cn; wangsy2006@126.com; 002470@nuist.edu.cn; 202212490688@nuist.edu.cn

Received Nov. 17, 2023; Revision accepted Feb. 26, 2024; Crosschecked Sept. 5, 2024

Abstract: Human emotions are intricate psychological phenomena that reflect an individual's current physiological and psychological state. Emotions have a pronounced influence on human behavior, cognition, communication, and decision-making. However, current emotion recognition methods often suffer from suboptimal performance and limited scalability in practical applications. To solve this problem, a novel electroencephalogram (EEG) emotion recognition network named VG-DOCoT is proposed, which is based on depthwise over-parameterized convolutional (DO-Conv), transformer, and variational automatic encoder-generative adversarial network (VAE-GAN) structures. Specifically, the differential entropy (DE) can be extracted from EEG signals to create mappings into the temporal, spatial, and frequency information in preprocessing. To enhance the training data, VAE-GAN is employed for data augmentation. A novel convolution module DO-Conv is used to replace the traditional convolution layer to improve the network. A transformer structure is introduced into the network framework to reveal the global dependencies from EEG signals. Using the proposed model, a binary classification on the DEAP dataset is carried out, which achieves an accuracy of 92.52% for arousal and 92.27% for valence. Next, a ternary classification is conducted on SEED, which classifies neutral, positive, and negative emotions; an impressive average prediction accuracy of 93.77% is obtained. The proposed method significantly improves the accuracy for EEG-based emotion recognition.

Key words: Emotion recognition; Electroencephalogram (EEG); Depthwise over-parameterized convolutional (DO-Conv); Transformer; Variational automatic encoder-generative adversarial network (VAE-GAN)
https://doi.org/10.1631/FITEE.2300781 **CLC number:** TP183

1 Introduction

Emotion recognition has attracted more and more attention in recent years as emotional computing, proposed by Picard (2000) continues to evolve. The main challenge in emotional computing is to enable computer systems to accurately process, identify, and understand the emotional information expressed by people, thus realizing human–computer interaction (HCI). Emotion is an incredibly intricate state of mind. Traditionally, emotions are identified by extracting emotional states from various physical behaviors and activities, such as facial expression, natural language, body posture, writing, and other signals. The judgement of emotional states based on physiological activities (physiological cues) has become a hot topic in affective computing. Some physiological studies have shown that there is a relationship between physiological processes and affective cognitive processes, although there is a debate on the occurrence order of the two processes (Li X et al., 2022). Therefore, computational psychophysiology relying on facial or voice information is considered an effective method in emotion recognition techniques, which are non-physiological cues.

‡ Corresponding author

* Project supported by the National Key Research and Development Program of China (No. 2022YFE0122700) and the National Natural Science Foundation of China (No. 61971230)

ORCID: Yanping ZHU, https://orcid.org/0000-0002-7366-1972

The signal feature does not change with the user's deliberate disguise, and it can be the most accurate reflection of an individual's emotional state. Due to high temporal resolution and highly accurate temporal information, electroencephalogram (EEG) allows researchers to observe rapid and dynamic changes in brain activity during emotional processes, revealing the temporal nature of emotions and the differences between diverse emotional states. Therefore, the utilization of EEG to identify emotional states is attractive and valuable.

Many efforts have been made to improve the accuracy of EEG-based emotion recognition. Feature signal extraction is one of the critical steps in achieving a high recognition rate. It is essential to process the raw EEG data in order to obtain a representative feature set. The preprocessing of EEG signals generally includes artifact removal (such as electrooculogram artifacts, electrocardiogram artifacts, and some power frequency artifacts), down sampling, filtering, and component decomposition (Yang BH et al., 2015). When data are incomplete or particularly poor due to technical or personal issues, it is necessary to discard them. In EEG-based emotional recognition, there are three kinds of signal representations: time domain, frequency domain, and time–frequency domain (Bernat et al., 2001; Lin et al., 2009; Sorkhabi, 2014; Zheng and Lu, 2015; Mohammadi et al., 2017; Liu YJ et al., 2018; Tang et al., 2019). Additionally, the selecting and building of brain channels' networks has been employed during the extraction of feature signals. In emotion recognition, an increasing adoption of nonlinear dynamic characteristics and chaotic characteristics in the nervous system has been witnessed (Stam, 2005), such as fractal dimension (Wang Q et al., 2011), Shannon entropy (Vijayan et al., 2015), sample entropy (Hu and Min, 2018), and differential entropy (DE) (Liu YS and Sourina, 2014; Yang Y et al., 2021). Among these features, DE has shown exceptional effectiveness and robustness, as validated in various related works (Zheng and Lu, 2015).

To further improve the recognition rate, it is critical to expand the limited dataset. Currently, generative adversarial networks (GANs) (Goodfellow et al., 2020) and variational autoencoders (Kingma and Welling, 2014) are widely used to augment the data during model training. Zhang QQ and Liu (2018) proposed a conditional deep convolutional GAN (cDCGAN), which resulted in an improvement in classification accuracy from 83% to 84%. Furthermore, Aznan et al. (2019) employed variational automatic encoder (VAE) to generate synthetic EEG signals to solve the problem of insufficient data and show the effectiveness of synthetic data in enhancing the classification performance.

Feature extraction and data enhancement need to extract low-level features from EEG, while the core model of emotional recognition primarily relying on machine learning or deep learning needs to obtain high-level features automatically. For instance, Wang XW et al. (2014) extracted two power spectrum features, two wavelet signs, and nonlinear dynamic features from EEG signals to investigate the relationship between EEG data and emotional states. They also proposed three dimensionality reduction methods and employed a linear support vector machine (SVM) classifier to classify two emotions. Jenke et al. (2014) extracted time–frequency domain features and channel combination features from multi-channel EEG. Lan et al. (2016) adopted SVM to build a recognition model based on statistical features and nonlinear features, comparing its performance on user data from the same period and different periods. Cheng et al. (2021) proposed a deep forest model named multi-grained cascade forest (gcForest). However, traditional machine learning methods have some limitations when dealing with EEG signals. They often struggle to handle the complexity of the data and have limited ability to learn hierarchical representations. These methods typically learn shallow representations, resulting in lower accuracy in executing complex tasks.

Deep learning has gained significant popularity in emotional computing due to its learning ability and outstanding performance in dealing with complex problems. Zheng and Lu (2015) used deep belief networks to classify EEG-based emotions according to manual features extracted from EEG signals. Because convolutional neural networks (CNNs) are very suitable for processing two-dimensional (2D) data and extracting joint information between channels, they have been widely used in multi-channel EEG-based emotion recognition. Yang YX et al. (2018) proposed a channel-frequency CNN (CFCNN) that utilizes entropy features in different frequency bands obtained from recursive quantization analysis. Tripathi et al.

(2017) extracted nine statistical features from EEG signals as the input of the CNN model, which finally exceeded the performance of mainstream methods. To better represent data, Li JP et al. (2018) organized different entropy features from different channels into 2D sparse graphs, preserving the topological information of electrode space, and inputted them into the CNN model for training. Salama et al. (2018) proposed a 3D-CNN model where multi-channel EEG was randomly arranged into frames, and adopted a data enhancement method by adding Gaussian white noise to the original signal. Chao and Dong (2021) proposed an advanced CNN designed with a univariate convolution layer and a multivariate convolution layer to deal with the 3D feature matrix for emotion recognition. Li C et al. (2022) proposed an efficient CNN and contrastive learning (ECNN-C) method for EEG-based emotion recognition.

Previous studies often overlooked the temporal and wave characteristics of EEG signals, despite the fact that subjects' specific emotions usually evolve with EEG fluctuations during experiments. To address this problem, researchers focused on context modeling of EEG signals to capture the long signal dependence. Li X et al. (2016) designed a convolutional recurrent neural network (CRNN) model that can decode the relationship between channels and a RNN module that can capture context information. Zhang T et al. (2019) designed a spatial–temporal hybrid model named STRNN, which used RNN modules to learn temporal and spatial dependencies in a multi-channel context. Zhang DL et al. (2018) introduced a model of cascade and parallel hybrid CNN and RNN, and the performance was always superior to that of the state-of-the-art (SOTA) method. Yang YL et al. (2018b) proposed a similar parallel model, but it incorporated a preprocessing method to remove the baseline signal before the stimulus test, resulting in a 32% accuracy improvement. Wei et al. (2020) suggested using integrated simple recursive unit (SRU) networks to learn different EEG rhythm feature sequences obtained by wavelet transform.

Attention mechanisms, primordially proposed by Bahdanau et al. (2015), have become a research hotspot in deep learning. Tao et al. (2023) introduced an attention mechanism into the above CNN-long short term memory (CNN-LSTM) model. They integrated channel attention and self-attention into CNN and RNN, respectively, to learn attention features between channels and within sequences. Lew et al. (2020) proposed a model based on gate recurrent unit-RNN (GRU-RNN) and introduced a regional attention layer to assign weight to increase or decrease the influence of different regions. The transformer model, proposed by Vaswani et al. (2017), used attention mechanisms instead of RNNs, and it has exhibited excellent performance in the field of natural language processing (NLP). Zhong et al. (2023) put forward a novel bi-hemispherical asymmetric attention network (Bi-AAN) that combined the converter structure with the asymmetry of emotional responses in the brain. In this way, differences in attention between the two hemispheres of the brain were simulated, and long-term dependencies between the brain's electrical sequences were mined, leading to more differentiated representations of emotions. Li SJ et al. (2022) proposed a personality-guided attention neural network that can use personality information to learn effective EEG representations for emotion recognition. Guo et al. (2022) developed a new neural network model (DCoT) with deep convolution and transformer encoders for EEG-based emotion recognition by exploring the dependence of emotion recognition on each EEG channel and visualizing the captured features. In summary, various factors, including EEG characteristics (time, frequency, and spatial), signal preprocessing methods, network structures, and data augmentation techniques, influence the overall recognition rate of the model. This paper aims to conduct research based on the above factors to improve the performance of EEG emotion recognition systems.

In this work, a novel emotion recognition model called VG-DOCoT is proposed to improve the recognition rate. The model combines the strengths of depthwise over-parameterized convolutional (DO-Conv) into transformer architectures, referred to as DO-CoT. It can effectively capture low-order features of frequency bands, correlate long-term dependencies, and explore global features.

Meanwhile, a data augmentation technique using VAE-GAN is incorporated into this model. First, we convert the raw EEG signal into graph data containing the frequency, space, and time information, and then send the data into the VAE-GAN model as input

to generate a high-quality enhancement EEG signal for the subsequent model learning. Next, these data are inputted into the DO-Conv structure to learn the frequency and spatial representation of the EEG signal. These features are finally used as inputs to the transformer module to capture the global dependence of EEG signals. The major contributions of this work are summarized as follows:

(1) An improved feature structure is proposed that contains the frequency, spatial, and temporal information of EEG. It can maintain various data features of EEG signals and obtain data information more comprehensively, enabling more effective and advantageous data processing in subsequent steps.

(2) To address the issue of limited data availability, VAE-GAN is used as a data augmentation technique within the processing model. Specifically, the available data are artificially partitioned for training and test sets, and VAE-GAN is subsequently applied to the training data. The objective is to enhance the diversity of the training data, thereby improving the model's performance and generalization ability, as well as its robustness and accuracy.

(3) A novel model called VG-DOCoT is proposed, which is designed for an EEG-based emotion recognition technique. In this model, the DO-Conv structure is used to obtain the local features between different channels, and the transformer structure is used to obtain the time dependence of EEG signals. Experimental results demonstrate that the proposed model achieves a better performance than the existing approaches in terms of recognition accuracy, showing its effectiveness in the field of EEG-based emotion recognition.

2 Methods and materials

2.1 Database

DEAP: DEAP is a multimodal database for human emotion analysis, provided by BALab at the Imperial Institute of Technology in Piedmont, Italy (Koelstra et al., 2012). Each of the 32 participants in the database watched 40 music videos with different emotional content. Each video was 1-min long. Physiological signals, such as EEG, were collected at a sampling rate of 512 Hz. While watching different videos, each participant's physiological data were divided into 40 segments based on their states. Each section contained a 3-s baseline state and a 60-s trial state. The dataset includes four emotional dimensions, arousal, valence, dominance, and liking, ranging from 1 to 9. Participants' arousal, valence, dominance, and liking scores are used as classification criteria. A score of 1 to 5 and a score of 6 to 9 are divided into two categories, marked with 0 and 1, respectively.

SEED: The SJTU affective EEG dataset (SEED) is a collection of EEG datasets provided by the BCMI Laboratory headed by professor Baoliang Lv (Zheng and Lu, 2015). SEED contained the electrical EEG of subjects when they were watching a movie clip. Film clips were carefully selected to evoke different types of emotions, including positive, negative, and neutral emotions. Based on both visual and auditory stimuli, the dataset collects EEG signals from 15 subjects through 15 emotionally labeled video clips (consisting of 5 positives, 5 neutrals, and 5 negatives). A subject watched 15 clips in one experiment to collect 15 clips of data, each of which contained 62 conductors of EEG time series data. After some time, the data of each subject were collected again in the same way, and the data were collected three times in total; this means that SEED has three sessions. Therefore, a total of 3×15 experiments are needed, where 15 is the number of subjects and 3 is the number of experiments conducted for each subject. Each group of experiments in this dataset includes $15\times62\times M$ EEG signals, where 15 is the number of emotional inductions in each group, 62 is the number of brain electrodes, and M is the number of samples for each clip.

Table 1 shows the format of the two datasets. "Data" and "Label" are the total dimensions of the data, and "data" and "label" are the dimensions of a single EEG channel.

Table 1 Dataset description

Dataset	Name	Shape	Content
DEAP	Data	32×40×32×8064	Subjects×videos×channels×data
	Label	32×40×1	Subjects×videos×label
SEED	Data	$3\times15\times15\times62\times M$	Sessions×subjects×videos×channels×data
	Label	3×15×15×1	Sessions×subjects×videos×label

2.2 Preprocessing and feature extraction

Due to the influence of various external factors, individual emotions can be highly variable. To address this issue, the EEG signal is preprocessed by applying a bandpass filter to eliminate power frequency interference. Specifically, a 0–45 Hz bandpass filter is utilized for the DEAP dataset, while a 0–75 Hz bandpass filter is applied for SEED. The DEAP dataset contains a baseline signal; to better capture the emotional state of the subjects, the baseline removal method is introduced. Since SEED does not have a baseline signal portion, baseline signal removal is not required. The formula for computing the average baseline signal is defined as

$$\text{BasicMean} = \sum_{1}^{3} \text{Basic}/3, \tag{1}$$

where BasicMean is the average baseline signal for the first three seconds and Basic is the baseline signal for each of the first three seconds. After computing the average baseline signal value, the emotional signal captured during the 60-s interval is partitioned into 60 segments using a sliding window of 1 s each. The value of each emotional segment is then subtracted from the average baseline signal value, resulting in a signal where the baseline had been effectively removed. These calibrated EEG emotional signals are subsequently treated as novel signals for the ensuing processing steps. The following step is feature extraction. We adopt the DE method, which has an outstanding performance on EEG signal recognition. Its mathematical definition is:

$$h(k) = \int_{k} f(k)\log(f(k))\,\mathrm{d}k, \tag{2}$$

where k is a random variable and $f(k)$ is the probability density function of k. The variability of EEG signals across emotional states is captured by DE, which is effective in distinguishing these states due to its sensitivity to the signal's rapid changes and essential for tracking the swift EEG fluctuations tied to emotions. Unlike average features, DE can detect quick signal variations without being affected by the signal's amplitude scale, offering robustness against varying experimental setups and device calibrations. Its computational simplicity also makes it ideal for real-time processing in practical applications. The DE signal processing workflow is illustrated in Fig. 1.

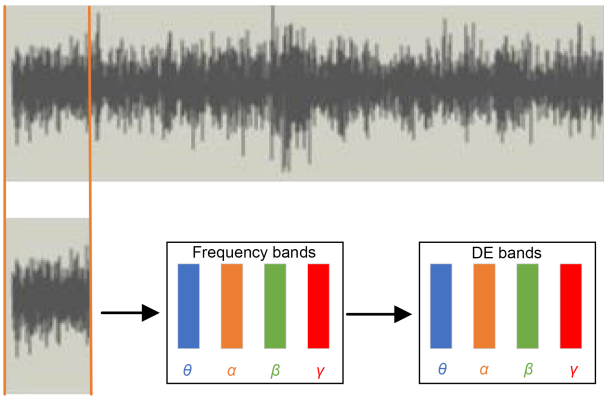

Fig. 1 Structure of differential entropy (DE) feature extraction

At first, the EEG emotional signals are partitioned into 120 non-overlapping segments by utilizing a sliding window of 0.5 s, and the 120 segments are labeled using the corresponding emotional tags from the original experiment. Next, each segment is subjected to a Butterworth filter, which decomposes it into four bands, namely, θ band (4–8 Hz), α band (8–14 Hz), β band (14–31 Hz), and γ band (31–45 Hz), as these bands have been demonstrated to account for the majority of emotional expression. In addition, for each 0.5 s fragment, the DE feature is extracted. Once obtaining the DE vector, normalization is performed on each vector to expedite the process and enhance the reliability of the results. To capitalize on the spatial attributes of EEG electrodes, the signals of each frequency band are combined according to the position of electrodes to obtain a 2D feature topology. The specific details of the electrode position diagram of the 10–20 system projected into the 2D feature topology diagram are shown in Figs. 2 and 3. Fig. 2 represents the mapping diagram of 32 EEG channels of system 10–20, while Fig. 3 illustrates the mapping of 62 EEG channels in system 10–20.

By superposing the 2-D topologies of the four frequency bands, the signals of each channel from a four-dimensional (4D) feature are defined as $\boldsymbol{X}_n \in \mathbb{R}^{h\times w\times d\times T}$, where h is the height of the 2D topology graph, w represents the width of the 2-D topology graph, d stands for the number of frequency bands, and T is the length of the segment partition.

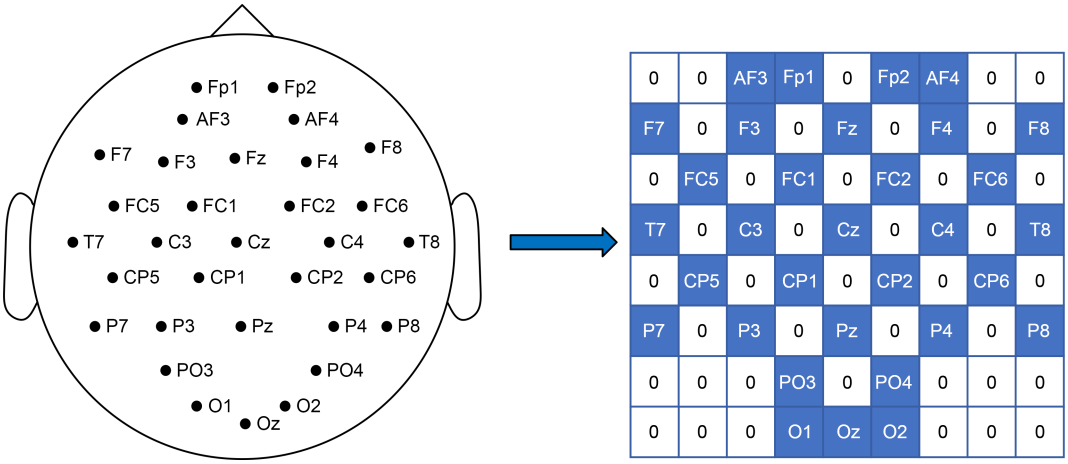

Fig. 2 10–20 system 32 channel spatial lead feature matrix

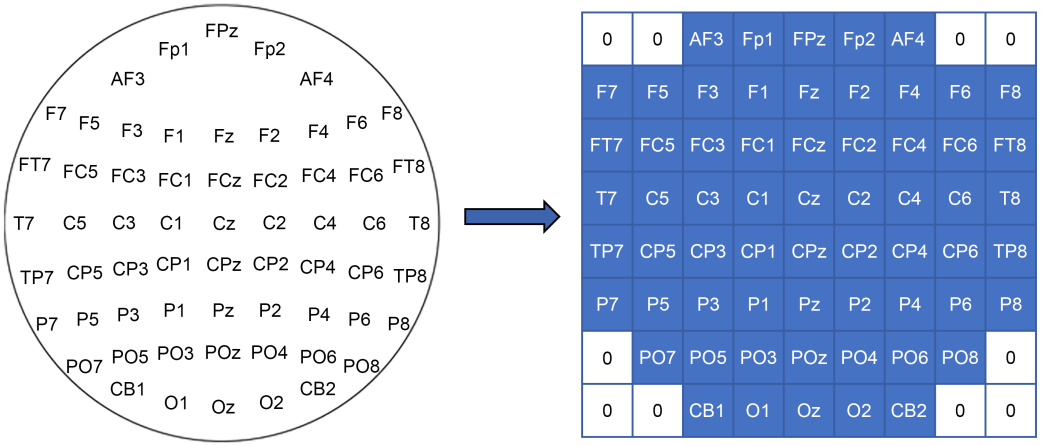

Fig. 3 10–20 system 62 channel spatial lead feature matrix

2.3 EEG emotional data augmentation model based on VAE-GAN

At present, the most popular data enhancement methods of machine learning include GAN (Goodfellow et al., 2020) and VAE (Kingma and Welling, 2014). VAE further performs variational processing on the traditional AE model to ensure that the output results of the encoder correspond to the mean and variance of the target distribution, thereby effectively improving classification performance. However, the quality of the data samples that VAE generate is unstable, and it is difficult to generate complex data. GAN continuously optimizes the data generated by itself through the generation network so that the discriminator network cannot distinguish between the real data and the generated data, thus generating high-quality fake data. These data lack effective control over the underlying space and may produce discontinuous or invalid data. By combining these two networks, the generative ability of GAN and the potential spatial control ability of VAE can be obtained, so as to generate higher-quality data for EEG emotion recognition.

This paper uses EEG signals $\boldsymbol{x}_{\text{real}}$ as the input of VAE. As shown in Fig. 4, input $\boldsymbol{x}_{\text{real}}$ into the encoder E to obtain a posteriori estimate of the data distribution

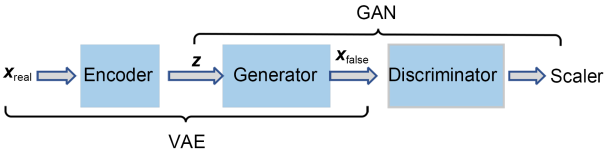

Fig. 4 Variational automatic encoder-generative adversarial network (VAE-GAN) model diagram

$q(\boldsymbol{z}|\boldsymbol{x}_{\text{real}})$. Then, the conditional distribution of the data $p(\boldsymbol{z}|\boldsymbol{x}_{\text{real}})$ is reconstructed under the constraint of the prior distribution $p(\boldsymbol{z})$ by inputting a low-dimensional latent variable $\boldsymbol{z}$ into the decoder G. $q(\boldsymbol{z}|\boldsymbol{x}_{\text{real}})$ and $p(\boldsymbol{z}|\boldsymbol{x}_{\text{real}})$ are usually represented by the following:

$$\boldsymbol{z}\sim E\left(\boldsymbol{x}_{\text{real}}\right)=q\left(\boldsymbol{z}|\boldsymbol{x}_{\text{real}}\right),\boldsymbol{x}'_{\text{real}}\sim G(\boldsymbol{z})=p\left(\boldsymbol{z}|\boldsymbol{x}_{\text{real}}\right), \quad (3)$$

where $E(\cdot)$ is the encoder, $G(\cdot)$ is a decoder, and $\boldsymbol{x}'_{\text{real}}$ is the sample of reconstruction. The potential vector $\boldsymbol{z}$ is composed of the combination of the mean value μ and the standard deviation σ of the output of the encoder E. The formula gives:

$$\boldsymbol{z}=\mu+\gamma\odot \mathrm{e}^{\sigma}, \quad (4)$$

where $\gamma\sim N(\boldsymbol{0},\boldsymbol{I})$ obeys a Gaussian distribution and $\odot$ represents an element multiplication operation; so it can be approximated that the potential vector $\boldsymbol{z}$ conforms to the Gaussian distribution, namely $\boldsymbol{z}\sim N\left(\boldsymbol{0},\boldsymbol{I}\right)$. The Kullback–Leibler (KL) divergence is introduced to optimize the parameters of the encoder. The KL divergence loss formula is as follows:

$$L_{\text{KL}}=\frac{1}{2}\left(\sum_{i=1}^{D}-1-\log\left(\sigma_i^2\right)+\sigma_i^2+\mu_i^2\right), \quad (5)$$

where L_{KL} is the calculation of KL divergence distance.

VAE not only uses KL loss to optimize the encoder, but also adopts reconstruction loss to optimize the decoder. The reconstruction loss can be formulated as

$$L_{\text{Rec}}=\frac{1}{2}\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_{\text{real}})}\left[\left\|G(\boldsymbol{z})-\boldsymbol{x}_{\text{real}}\right\|^2\right], \quad (6)$$

where $\mathbb{E}$ means the expected value of the corresponding distribution. The formula calculates the square of the Euclidean distance between the real data and the synthetic data.

Therefore, the total loss VAE is expressed as

$$\begin{aligned}L_{\text{VAE}}&=L_{\text{KL}}+L_{\text{Rec}}\\&=\frac{1}{2}\left(\sum_{i=1}^{D}-1-\log\left(\sigma_i^2\right)+\sigma_i^2+\mu_i^2\right)\\&\quad+\frac{1}{2}\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_{\text{real}})}\left[\left\|G(\boldsymbol{z})-\boldsymbol{x}_{\text{real}}\right\|^2\right].\end{aligned} \quad (7)$$

Since data samples generated by traditional VAEs can be less consistent in quality, especially when it comes to reproducing detail and maintaining high resolution, the introduction of GANs into models can learn complementary information to improve the quality of generated samples. A GAN consists of a generator (G) which is responsible for generating fake data that are as close as possible to the real data, and a discriminator (D) which attempts to distinguish between the generated fake data and the real data samples. In this structure, the discriminator D not only gives priority to samples from real data and gives them higher weights, but also facilitates the learning of the generator G through this process, making it produce more realistic data. The loss function of the discriminator in GAN is expressed as

$$\begin{aligned}L_D&=\frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{\text{real}}\sim p_{\text{data}}}\left[\left(D(\boldsymbol{x}_{\text{real}})-1\right)^2\right]\\&\quad+\frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}\left[D\left(G(\boldsymbol{z})\right)^2\right],\end{aligned} \quad (8)$$

where $D(\boldsymbol{x}_{\text{real}})$ is the output of the discriminator against the real samples, $G(\boldsymbol{z})$ is the sample generated by the generator through noise $\boldsymbol{z}$, and $p_{\boldsymbol{z}}$ is the input noise distribution of the generator.

VAE-GAN will be pitted against G and D:

$$\min_{D,\text{VAE}} L\left(\text{VAE},D\right)=L_{\text{VAE}}+L_D. \quad (9)$$

After training the entire model, the generator produces high-quality samples through Gaussian noise. In this study, the diagrams of DE features processed earlier are chosen as the input for the data enhancement model. The real sample input is split into a training set and a test set. For the training set, the VAE-GAN model is utilized to expand the data of various emotional labels. Specifically, data enhancement is carried out by doubling the original size on DEAP dataset and SEED. The real sample training set and the expanded fake sample training set are merged and utilized as new training sets, while the real sample test set is used for the classification task.

2.4 Architecture of the DO-CoT model

In this paper, we propose the DO-CoT model to provide an effective method with which to simultaneously capture the time, frequency, and spatial

information of EEG signals and complete the emotion recognition task. The input of the model is $\boldsymbol{X}_n \in \mathbb{R}^{h\times w\times d\times T}$. In this subsection, we combine a CNN and transformer model. DO-CoT's basic components include a DO-Conv layer, location embedding, a transformer's encoder layer, and a linear layer. The encoder layer of a transformer includes multiple attention layers and feedforward layers, as well as residual connection structures. An overview of the DO-CoT model is illustrated in Fig. 5.

2.4.1 DO-Conv layer

The 4D feature structure obtained earlier is utilized in this paper to extract spatial and frequency information using a CNN. Based on Yang YL et al. (2018a)'s CNN, the structure of our CNN is improved by replacing the part of the traditional convolutional layer with DO-Conv (Cao et al., 2022) and adding a maxpooling layer to prevent overfitting.

Given the input feature map, it is processed using the conventional convolutional layer in a sliding window fashion. In each window, there is a set of convolution kernels applied to the corresponding size of the feature map patch. This patch is called $\boldsymbol{P} \in \mathbb{R}^{(M\times N)\times C_{\text{in}}}$, where $M \times N$ is the size of the convolution kernel and also the size of the patch, and C_{in} is the number of channels in the input feature map. DO-Conv is a composition of depthwise convolution with trainable kernel $\boldsymbol{D} \in \mathbb{R}^{(M\times N)\times D_{\text{mul}}\times C_{\text{in}}}$ and a conventional convolution with trainable kernel $\boldsymbol{W} \in \mathbb{R}^{C_{\text{out}}\times D_{\text{mul}}\times C_{\text{in}}}$, where $D_{\text{mul}}=M\times N$ and C_{out} is the number of channels in the output feature map. The output of DO-Conv is $\boldsymbol{O}=\boldsymbol{W}*(\boldsymbol{D}\circ\boldsymbol{P})$, where $\circ$ is the dot product between each horizontal section of $\boldsymbol{D}$ and $\boldsymbol{P}$, and $*$ is the dot product between each vertical column of $\boldsymbol{W}$ and the corresponding channel element of $\boldsymbol{P}$. Cao et al. (2022) proved that the DO-Conv layer can improve the performance of many tasks by replacing the traditional convolutional layer. Moreover, the computational complexity of inference is not increased while the performance gain is introduced.

The proposed CNN structure contains four convolutional layers, a maxpooling layer, a flatten layer, and a full connection layer. The convolution kernel size of the first DO-Conv layer is 5×5 with 64 feature maps. The second DO-Conv layer has a convolution kernel size of 4×4 with 128 feature maps. The third DO-Conv layer has a convolution kernel size of 4×4 with 256 feature maps. The fourth convolution layer has a convolution kernel size of 1×1 with 64 feature maps. To maintain the size of the image after convolution, a step size of 1 and zero padding are used in all convolutional layers. A rectified linear unit activation function is applied to the layers. A maxpooling layer is incorporated after the fourth convolutional layer to avoid overfitting, improve network robustness, and minimize parameters. Considering the small size of the 2D feature topology, only one pooling layer is used. After pooling, a flatten layer converts the resulting 4×4×64 vector into a 1024-dimensional vector. Subsequently, a full connection layer reduces the dimension of the vector to 512-dimensional vector, and it is sent to the transformer. The final output of DO-Conv is $\boldsymbol{Q}_n=(\boldsymbol{q}_1, \boldsymbol{q}_2,\cdots,\boldsymbol{q}_T)$, where $\boldsymbol{q}_t \in \mathbb{R}^{512}$ and $t=1, 2, \cdots, T$. The novel DO-Conv structure is illustrated in Fig. 6.

2.4.2 Transformer model

We input the features $\boldsymbol{T}_n \in \mathbb{R}^{\text{batchsize}\times 6\times 512}$ obtained through the CNN into the transformer. Six is the sequence number for better transferring into transformer;

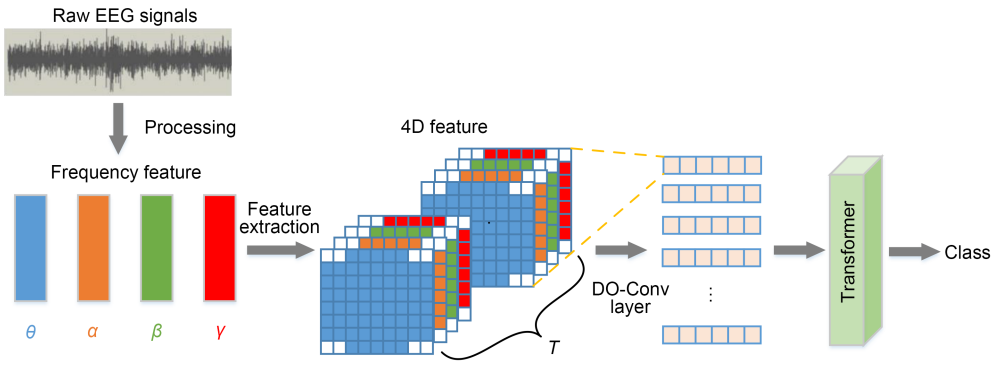

Fig. 5 Structure of the proposed DO-CoT model for emotion recognition

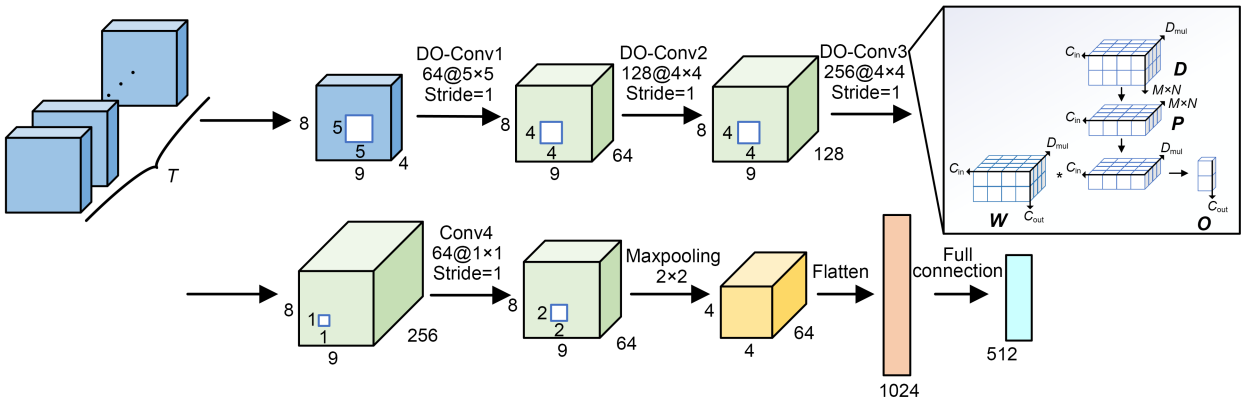

Fig. 6 Structure of DO-Conv model

512 is the characteristic information contained in each sequence. The transformer network (Vaswani et al., 2017) is illustrated in Fig. 7.

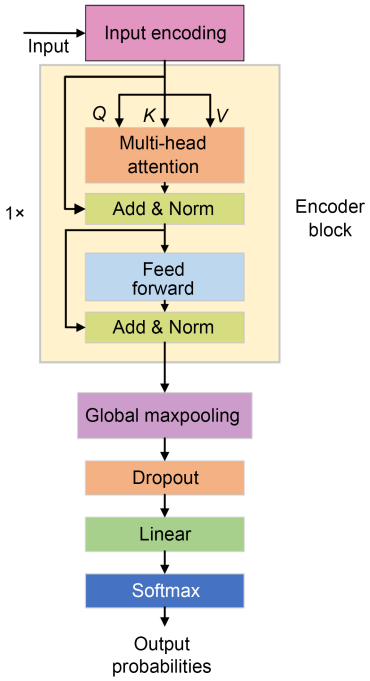

Fig. 7 Structure of the transformer model

First, position encoding is used to embed the position information of the data to import the input orders of feature vectors of different channels:

$$\text{input} = \text{input} + \text{positional}_{\text{encoding}}. \tag{10}$$

Among them, the input is $\boldsymbol{Q}_n \in \mathbb{R}^{512\times T}$, and the $\text{positional}_{\text{encoding}}$ is $\boldsymbol{P}_n \in \mathbb{R}^{512\times T}$. The dimensions are consistent to make it easy to add. The formulations of the position embeddings are based on sine and cosine functions of different frequencies, which can be expressed as

$$\text{PE}(\text{pos}, i) = \begin{cases} \sin\left(\dfrac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right), & i = 0, 2, \cdots, \\ \cos\left(\dfrac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right), & i = 1, 3, \cdots, \end{cases} \tag{11}$$

where d_{model} represents the number of input sequences, pos is the sequence order, and i indicates the dimension of each input sequence. Therefore, sinusoids with wavelengths ranging from 2π to $10\,000\times 2\pi$ are assigned to each dimension of the positional code. The matrix after position encoding is $\boldsymbol{X}\in\mathbb{R}^{512\times T}$.

The encoder layer contains a multi-head self-attention (MSA) layer and a feed forward neural network (FFN). The MSA layer integrates all the sequences of relevant channels into the sequence processing by transforming the input matrix $\boldsymbol{X}$ linearly into $\boldsymbol{QKV}$ space, i.e., queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$, where $\boldsymbol{Q}\in\mathbb{R}^{512\times d_{\text{q}}}$, $\boldsymbol{K}\in\mathbb{R}^{512\times d_{\text{k}}}$, $\boldsymbol{V}\in\mathbb{R}^{512\times d_{\text{v}}}$.

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^{\text{Q}}, \\ \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^{\text{K}}, \\ \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^{\text{V}}, \end{cases} \tag{12}$$

where the weighting matrices are $\boldsymbol{W}^{\text{Q}}\in\mathbb{R}^{T\times d_{\text{q}}}$, $\boldsymbol{W}^{\text{K}}\in\mathbb{R}^{T\times d_{\text{k}}}$, and $\boldsymbol{W}^{\text{V}}\in\mathbb{R}^{T\times d_{\text{v}}}$. The number of MSA layers (H) is set to 8. Matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ are obtained by using the learnable parameters $\boldsymbol{W}^{\text{Q}}$, $\boldsymbol{W}^{\text{K}}$, and $\boldsymbol{W}^{\text{V}}$ of H different groups, each projection using a different weighting matrix.

The essence of the projection of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ based on MSA layers is to project the vector to different representation subspaces. The specific formulas of MSA layers are as follows:

$$\text{Attention}\left(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\right)=\text{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}_i, \quad (13)$$

$$\text{head}_i=\text{Attention}\left(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\right), \quad (14)$$

$$\text{Multihead}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right)=\text{Concat}\left(\text{head}_1, \text{head}_2, \cdots, \text{head}_H\right)\boldsymbol{W}^o, \quad (15)$$

where softmax$(\cdot)$ represents softmax mapping. The internal parameters of the weighting matrices $\boldsymbol{W}_i^{\mathrm{Q}}$, $\boldsymbol{W}_i^{\mathrm{K}}$, $\boldsymbol{W}_i^{\mathrm{V}}$, and $\boldsymbol{W}^o$ can be learned and trained. After random initialization, they are constantly updated and corrected during backpropagation or feedback check. After determining the loss function of the network loss and setting the learning ratio η, the new weighting matrix is:

$$\boldsymbol{W}_{\text{new}} = \boldsymbol{W} - \eta\frac{\partial \text{loss}}{\partial \boldsymbol{W}}, \quad (16)$$

where the loss is the loss function of the network, and the cross-entropy function is selected in this paper.

In the process of coding, an output matrix $\boldsymbol{M}$ is generated and added with input matrix $\boldsymbol{I}$ by the pattern of the residual neural network, and then a new matrix $\boldsymbol{P}$ is generated through layer normalization. This new matrix $\boldsymbol{P}$ will generate a new matrix $\boldsymbol{Q}$ through the feedforward layer, which will also be added to matrix $\boldsymbol{P}$ according to the residual neural network mode and normalized through the layer. Our model has a global maxpooling layer, a dropout layer, and a linear layer. Finally, the softmax function is used to calculate $\boldsymbol{X}_{\text{class}}$, which is the classification probability.

3 Experiments

We perform subject-dependent experiments on the DEAP dataset and SEED to demonstrate the emotion recognition ability of the proposed model. The transformer modules in the proposed model and comparison model are trained from scratch. First, we conduct several experiments on the network hyperparameters to obtain the most suitable value for this emotion recognition task. Then, we run a large number of experiments on different datasets to test the performance of the proposed network model with or without data augmentation. Finally, we compare and analyze our method with other different networks. The effectiveness is demonstrated on the DEAP dataset and SEED using the five-fold cross-validation method. The training and test data are obtained from the same subject and experimental trial. According to Section 2.2, the EEG signal is filtered, data normalized, baseline removed, and DE feature extracted successively. We turn the EEG timing signals of each channel into a 2D topology according to the EEG electrode distribution diagram of the 10–20 system. In this paper, we set the height of the 2D feature map h=8, the width of the 2D feature map w=9, and the number of EEG frequency bands d=4. For the DEAP dataset, each EEG channel is divided into 120 samples with a time window of 0.5 s. Using 32 channels in each subject, the subject's EEG is divided into 3840 samples. As there are 32 subjects, the total sample size of the entire dataset is 122 880. Then, we administer dichotomous emotion tasks on arousal and valence in the DEAP dataset. As for SEED, the samples are also divided using a sliding time window of 0.5 s. There are 6788 samples for each of the 15 subjects, and each subject performs three experiments, giving a total of 305 460 samples. In this work, we take accuracy and the confusion matrix as test metrics.

3.1 Effect of hyperparameters

This subsection primarily examines the impact of various critical factors on the model, such as the learning rate of the optimizer for training the deep neural network and the batch size of the input data. To determine the optimal learning rate for this model, we hold the batch size at 64 and use all frequency bands in combination, and then train the model using different learning rates.

Table 2 presents the performance of the network under different learning rates for arousal and valence prediction in the DEAP dataset and three emotional states in SEED. The results show that when the learning rate is set to 0.0001, the model has the best effect in three datasets, namely, arousal prediction and valence prediction in the DEAP dataset and SEED. The

Table 2 Accuracy of the model under different learning rates

Learning rate	Accuracy (%)		
	DEAP_arousal	DEAP_valence	SEED
0.001	88.20±3.61	87.98±3.76	86.21±3.99
0.0001	91.51±1.89	91.13±2.68	91.32±0.48
0.000 01	89.37±3.28	88.97±3.58	89.78±0.63
0.000 001	68.97±2.92	62.63±2.63	73.86±1.21

classification accuracy of these three datasets is 91.51%, 91.13%, and 91.32%, respectively, and the standard deviation is lower, indicating that the results are more stable and reliable compared to other learning rates. Although the standard deviation is lower when the learning rate is 0.000 001, the accuracy of the model is too low, making it unsuitable for this task.

Since the optimal model performance is achieved at a learning rate of 0.0001 in the three experiments, we fix the learning rate at 0.0001 and change the batch size of the input data to find the most suitable batch size for our model. The experimental results of different batch sizes are shown in Table 3.

Table 3 Accuracy of the model under different batch sizes

Batch size	Accuracy (%)		
	DEAP_arousal	DEAP_valence	SEED
16	86.42±2.35	85.98±2.76	91.32±0.48
32	87.03±2.98	85.97±2.58	90.59±0.36
64	91.51±1.89	91.13±2.68	89.46±0.98
128	88.48±2.99	87.63±2.38	88.18±0.76

Table 3 illustrates that the best optimal performance of the model for arousal and valence prediction in the DEAP dataset is achieved when the batch size is set to 64. In contrast, for the three-classification task of SEED, the network achieves the best performance with a batch size of 16, reaching an accuracy of 91.32% and a standard deviation of 0.48%.

3.2 Performance of DO-CoT

The following describes the overall performance of the proposed model on the DEAP dataset and SEED.

As can be seen in Fig. 8, the average prediction accuracy in terms of arousal of the DEAP dataset is 91.51%. The prediction accuracy of 19 subjects is higher than the average accuracy. Subject 22 has the lowest accuracy of 75.5% in this experiment, which is possibly due to inaccurate reporting of real emotions during data collection. Fig. 9 shows the prediction accuracy of valence in the DEAP dataset, with an average accuracy of 91.13%. For 16 of the 32 subjects, the prediction accuracy is higher than the average predicted. As can be seen from Fig. 10, the average prediction accuracy of SEED in this model is 91.32%; more than half of the subjects are predicted better than average accuracy.

3.3 Effect of the DO-Conv module

To demonstrate the effectiveness of our proposed DO-Conv+transformer (DO-CoT) in enhancing emotion recognition accuracy, we conduct ablation experiments to compare it with the conventional CNN+transformer (CoT). We train two models using the same hyperparameters and training settings: one with the DO-CoT module and the other with the conventional CoT module. Both models are trained and tested on the same training set. As can be seen in Table 4, after utilizing the DO-Conv module, the accuracy of

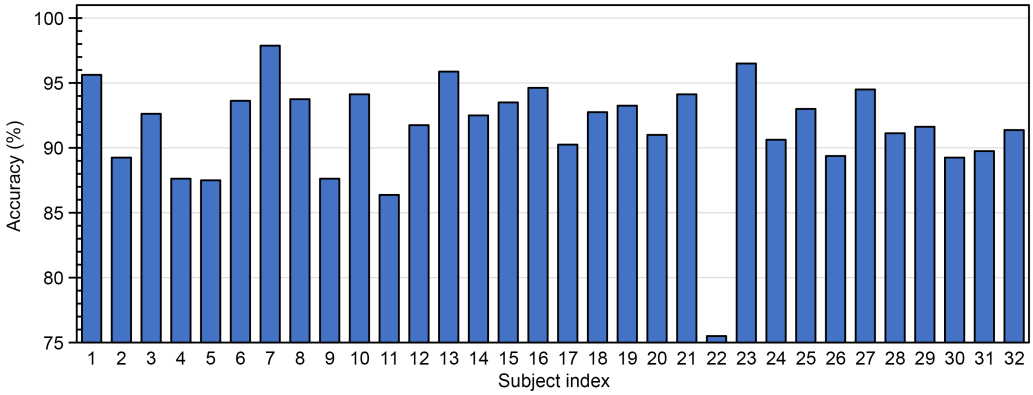

Fig. 8 Accuracy of arousal on the DEAP dataset

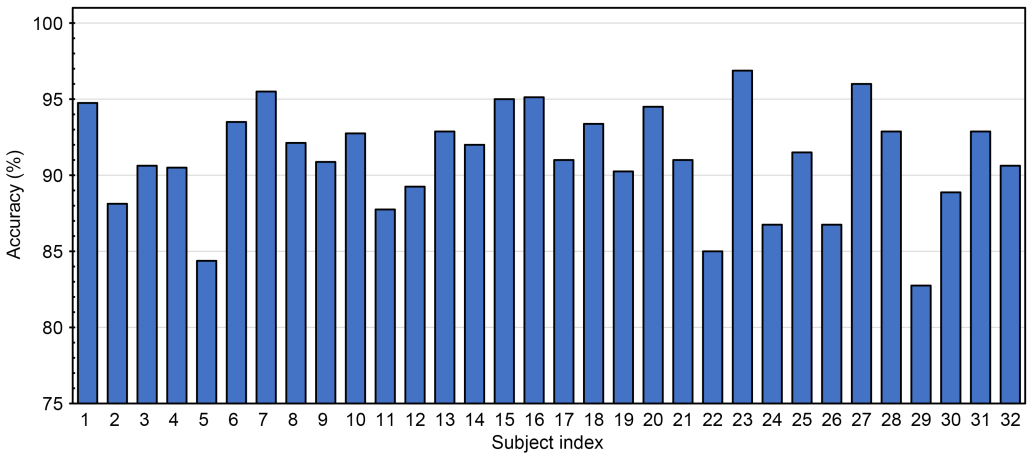

Fig. 9 Accuracy of valence on the DEAP dataset

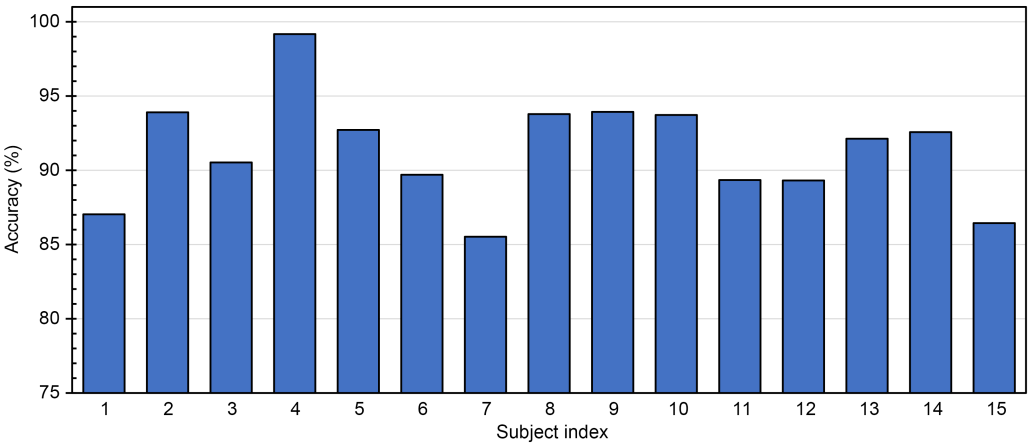

Fig. 10 Accuracy on SEED

Table 4 Accuracy comparison between CoT and DO-CoT

Model	Accuracy (%)		
	DEAP_arousal	DEAP_valence	SEED
CoT	91.39±1.92	91.09±2.15	90.88±0.48
DO-CoT	91.51±1.89	91.13±2.11	91.32±0.41

our model increases by 0.12%, 0.04%, and 0.44% in the arousal and valence of the DEAP dataset and SEED, respectively, and the standard deviation also decreases. The results indicate that our proposed DO-Conv module is a more effective network combination for enhancing emotion recognition accuracy than the traditional CNN module.

3.4 Effect of the transformer module

The transformer network is able to model the inter-dependencies between different regions of the input image, in addition to capturing sequential information. In addition, the parallel processing mechanism of the transformer model makes it possible to increase the performance of the model with lower computational costs. We compare the performance of the DO-CoT module with that of the DO-Conv model alone in our experiments. The DO-Conv model consists of only CNNs, which are commonly used for feature extraction in image processing. As can be seen in Table 5, the addition of the transformer network to the DO-Conv model significantly improves the emotion recognition accuracy, by 0.82% on arousal, 0.54% on valence, and 0.36% on SEED. This improvement highlights the effectiveness of the transformer network in capturing long-term dependencies in the input data and refining the features extracted by the DO-Conv.

Table 5 Accuracy comparison between DO-Conv and DO-CoT

Model	Accuracy (%)		
	DEAP_arousal	DEAP_valence	SEED
DO-Conv	90.69±1.87	90.59±2.12	90.96±1.62
DO-CoT	91.51±1.89	91.13±2.11	91.32±0.41

3.5 Effect of the VAE-GAN module

In addition to the proposed DO-CoT module, we utilize a novel data augmentation technique, VAE-GAN, to improve the model's generalization performance. We apply VAE-GAN to generate additional EEG data and use them to augment the training set, increasing the overall size of the dataset by a factor of two.

From Figs. 11–13, it is evident that the utilization of the VAE-GAN model to expand the training set leads to an average valence prediction accuracy of 92.52% on the DEAP dataset, which is 1.01% higher than that achieved without the VAE-GAN model. However, subject 22 has an accuracy of 74.5% which is not included in the table above due to its low accuracy. The average valence prediction accuracy in the DEAP dataset reaches 92.27%, which is 1.14% higher than that of the previous model. On SEED, the average prediction accuracy reaches 93.77% with the VAE-GAN model, which is 2.45% higher than that of the previous model. These results demonstrate the suitability of the VAE-GAN model for expanding EEG datasets.

To test the effect of the VAE-GAN module, we conduct a comparative experiment; the results are shown in Table 6.

Our experiments are conducted on three datasets, with the sample size progressively increasing from 2400 to 14400. It is observed that there is a consistent upward trend in accuracy for all datasets as the volume of samples increases. However, this improvement comes at the expense of increased computational time. To strike an optimal balance between accuracy and time, this study selects to augment the sample quantity by 4800. The decision is a balance between accuracy gains and computational resource constraints. It is based on the observation that accuracy improvements tend to plateau beyond this point, implying minimal benefits relative to the increased computational demand.

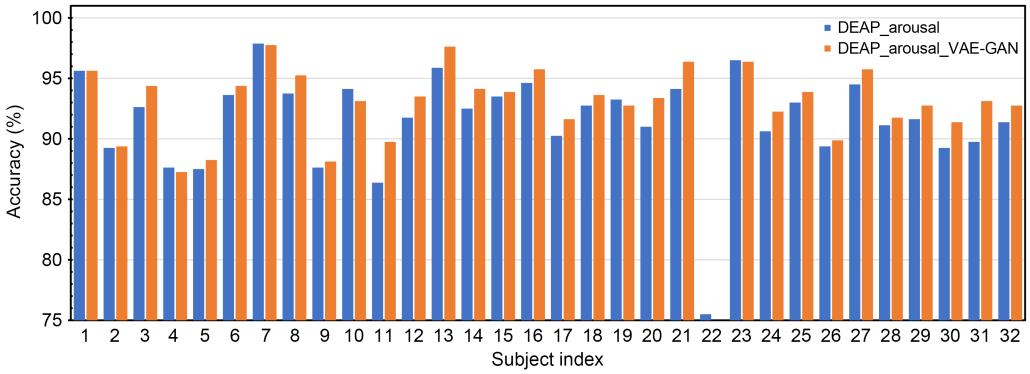

Fig. 11 Accuracy of arousal comparison on the DEAP dataset

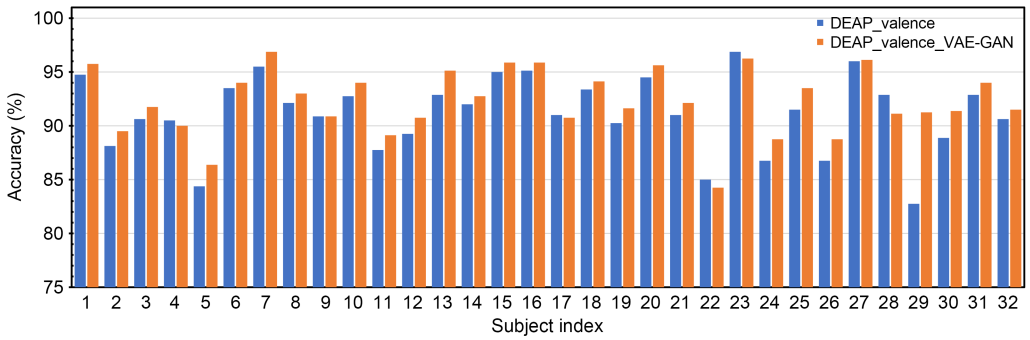

Fig. 12 Accuracy of valence comparison on the DEAP dataset

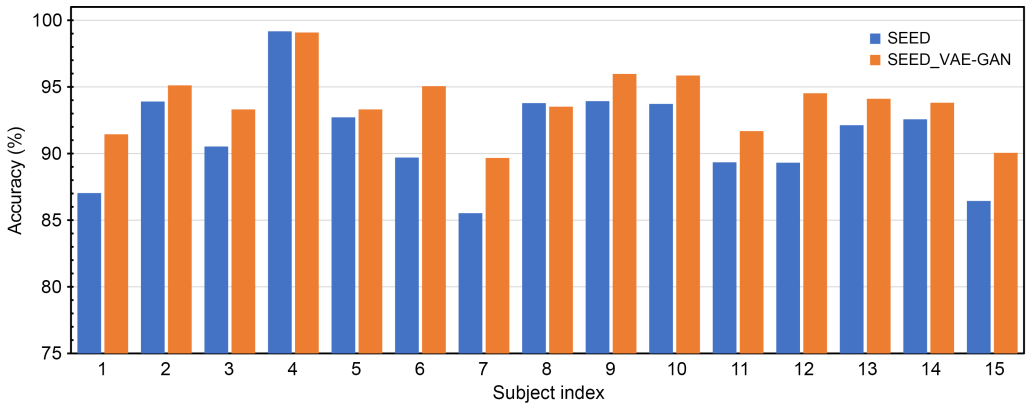

Fig. 13 Accuracy comparison on SEED

Table 6 Performance comparison of different sample sizes of VAE-GAN

Dataset	Accuracy (%)			
	Sample size=2400	4800	9600	14400
DEAP (arousal)	91.81±2.13	92.52±2.27	92.79±2.13	93.01±1.98
DEAP (valence)	91.51±2.18	92.27±2.55	92.56±2.11	92.87±2.08
SEED	92.56±0.63	93.77±0.62	94.01±0.84	94.23±0.77

3.6 Confusion matrix

To better analyze our model's recognition ability under different emotional states, this subsection describes the confusion matrices of our proposed networks on the DEAP dataset and SEED and the confusion matrix with the VAE-GAN structure. As can be seen from Figs. 14–16, the prediction accuracy of the low arousal and high arousal in the DEAP dataset is 88.26% and 93.90%, respectively, and the prediction accuracy of high arousal is higher. The reason might be attributed to more distinct physiological markers or more consistent patterns in EEG signals associated with high arousal. High arousal emotions, such as excitement or anger, could produce more pronounced

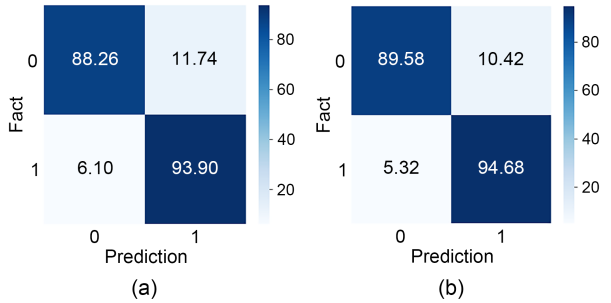

Fig. 14 Confusion matrix of arousal on the DEAP dataset (a) and confusion matrix of arousal with VAE-GAN on the DEAP dataset (b)

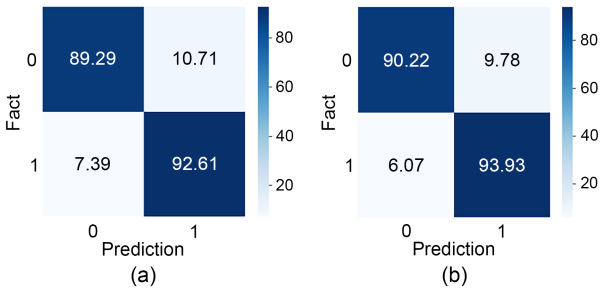

Fig. 15 Confusion matrix of valence on the DEAP dataset (a) and confusion matrix of valence with VAE-GAN on the DEAP dataset (b)

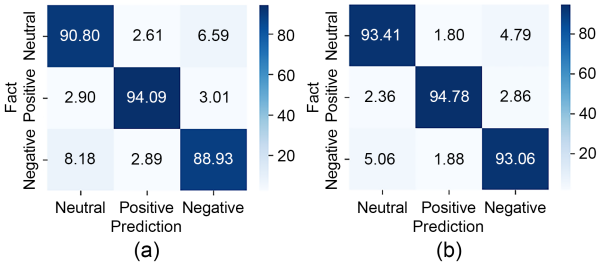

Fig. 16 Confusion matrix on SEED (a) and confusion matrix with VAE-GAN on SEED (b)

EEG patterns than low arousal emotions like calmness or relaxation. The prediction accuracy of the low and high valence are 89.29% and 92.61%, respectively, and the prediction accuracy of the high valence is higher as well. This could suggest that positive emotions

(high valence) like happiness or joy generate more distinguishable EEG signals than negative emotions (low valence), such as sadness or disgust. In SEED, the prediction accuracy of neutral, positive, and negative emotions are 90.80%, 94.09%, and 88.93%, respectively. Positive emotions are the most easily detected, while negative emotions and neutral emotions are easily confused. This indicates that the EEG patterns of positive emotions are more distinct or consistent, making them easier to detect. The confusion between negative and neutral emotions suggests similarities in their EEG signatures. This might be due to the subtlety of emotional states or a potential overlap in the EEG patterns of these emotions. In summary, the results demonstrate the model's promising capabilities in emotion recognition using EEG data, with specific strengths in detecting high arousal and positive emotional states. The addition of the VAE-GAN structure for data augmentation significantly contributes to the model's performance, highlighting the importance of diverse and comprehensive training datasets in emotion recognition tasks.

3.7 Method comparison

In recent years, many research groups have conducted important research on emotion recognition using EEG based on the DEAP dataset and SEED. As we can see from Table 7, these models include deep CNN, 3D-CNN, a model that combines CNN and RNN, a graph convolution neural network-LSTM (GCNN-LSTM) hybrid model, hierarchical convolution neural network (HCNN), and a dynamical GNN. From the DEAP dataset, we can see that, when the network uses only the simple CNN model, the performance is not very high, such as with deep-CNN and 3D-CNN. The proposed model demonstrates a significant improvement of 10.86% and 19.16% over the deep and convolutional neural networks model suggested by Tripathi et al. (2017). In comparison to the 3D-CNN model proposed by Salama et al. (2018), our model improves valence and arousal accuracy by 4.83% and 4.03%, respectively. The accuracy is enhanced by combining the RNN with the CNN model since EEG signals contain rich temporal information. But the proposed model still performs better than these models. Compared to the CNN-RNN fusion model presented by Zhang DL et al. (2018), our proposed model achieves an improvement of 1.47% and 1.49%. Additionally, compared to the GCNN-LSTM hybrid model suggested by Yin et al. (2021), our model demonstrates an improvement of 1.82% and 1.91% in the valence degree and arousal degree of the DEAP dataset, respectively. From SEED, our proposed method outperforms HCNN (Li JP et al., 2018) and dynamical GNN (Song et al., 2020) by 5.17% and 3.37%, respectively. Therefore, we can conclude that our proposed method, which combines CNN and a transformer while considering the frequency, space, and time information of EEG, is more suitable for emotion recognition tasks. Furthermore, we find that the addition of the VAE-GAN structure results in improved model performance, as more training data are utilized. Thus, we anticipate that our proposed model will achieve better results in practical applications since EEG data in such scenarios are more extensive.

4 Discussion

The above analysis shows that our VG-DOCoT structure performs better than other methods. This section discusses several notable issues. Compared

Table 7 Performances of the compared methods

Reference	Method	Dataset	Accuracy
Tripathi et al. (2017)	Deep and convolutional neural networks	DEAP	Valence: 81.41%; Arousal: 73.36%
Salama et al. (2018)	A 3D-CNN	DEAP	Valence: 87.44%; Arousal: 88.49%
Zhang DL et al. (2018)	Integrate CNN and RNN	DEAP	Valence: 90.80%; Arousal: 91.03%
Yin et al. (2021)	A GCNN-LSTM hybrid model	DEAP	Valence: 90.45%; Arousal: 90.60%
Ours	DO-Conv and transformer with VAE-GAN	DEAP	Valence: 92.27%; Arousal: 92.52%
Li JP et al. (2018)	HCNN	SEED	88.60%
Song et al. (2020)	A dynamical GNN	SEED	90.40%
Ours	DO-Conv and transformer with VAE-GAN	SEED	93.77%

with pure CNN networks or RNN networks, the proposed network in this paper achieves a better performance, because we take frequency and time information into account. Additionally, compared with an ordinary network that combines CNN and a transformer in parallel, our network has a higher emotion recognition rate. The reason is that a deep integration of DO-Conv and the transformer is utilized. The proposed VG-DOCoT first uses CNN to extract spatial features and frequency features from the feature containing EEG information, then uses the transformer network to extract time features from the output of the CNN module, and finally uses the transformer module output for classification. The spatial topological information of electrodes is taken into account when extracting EEG features in the proposed structure, thus achieving a better effect. Therefore, our proposed method can find more emotional information for emotion classification than other methods and achieve better results. In deep learning, it is generally required that the number of samples should be sufficient; the trained model has a good effect and strong generalization ability after the VAE-GAN structure is applied for data enhancement, and the network showed higher accuracy than before. This is also related to the fact that the transformer network is better at handling large amounts of data. Moreover, this paper also uses some data preprocessing methods to achieve a better effect. In the future, we will explore high-performance networks that do not rely on data preprocessing and data enhancement. Furthermore, we plan to significantly broaden the scope of our datasets. By incorporating EEG data from a diverse array of demographic groups, including various populations and cultural backgrounds, we aspire to achieve a more holistic and inclusive evaluation of our emotion recognition system. This expansion is crucial, as emotional expression and EEG patterns can vary significantly across different cultures and populations. A more diverse dataset will allow us to test the universality and adaptability of our approach, ensuring that our models are not only accurate but also equitable and inclusive. Cross-subject EEG emotion recognition is one of the hot topics in current research, which has a strong practical application value; we will continue to deeply research in this direction. Our research will continue to push the boundaries in this field, seeking ways to enhance the accuracy, efficiency, and applicability of EEG-based emotion recognition systems. The ultimate goal is to develop models that are not only technically advanced but also widely accessible and useful in real-world scenarios.

5 Conclusions

In conclusion, this paper introduces a groundbreaking emotion recognition method that synergizes a 4D feature structure with advanced techniques such as DO-Conv, transformer, and VAE-GAN. The core of our approach lies in the novel DO-Conv structure, designed to intricately extract both unique and interrelated EEG channel information. Complementing this, the transformer structure adeptly captures global EEG signal dependencies, offering a comprehensive analysis of emotional states. Our method significantly advances the field of emotion recognition by effectively utilizing the rich information inherent in EEG signals. This is particularly evident in the remarkable performance improvements seen with the DO-Conv and transformer networks. A key innovation of our approach is the use of data enhancement techniques, which address the prevalent challenge of limited EEG data, thereby bolstering the model's performance and reliability. The empirical results underscore the method's efficacy: on the DEAP dataset, we achieved an average accuracy of 92.27% in valence and 92.52% in arousal classification tasks. Moreover, on SEED's three-classification task, our method attained an impressive average accuracy of 93.77%. These outcomes not only demonstrate superior performance over existing methods but also highlight our approach's practical applicability and robustness. This research contributes to the emotion recognition field by offering a scalable, efficient, and highly accurate method. It opens new avenues for understanding emotional states through EEG analysis and sets a precedent for future work in this domain. Future research could explore the adaptability of this method to other neural signal types and emotional datasets, potentially broadening its applicability and furthering our understanding of the intricate relationship between neural patterns and emotional states.

Contributors

Yanping ZHU and Lei HUANG designed the research. Lei HUANG processed the data and analyzed the experimental results. Lei HUANG and Jixin CHEN made the charts. Lei HUANG and Yanping ZHU drafted the paper. Jixin CHEN and Jianan CHEN helped compile the paper. Yanping ZHU, Shenyun WANG, and Fayu WAN revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Aznan NKN, Atapour-Abarghouei A, Bonner S, et al., 2019. Simulating brain signals: creating synthetic EEG data via neural-based generative models for improved SSVEP classification. Int Joint Conf on Neural Networks, p.1-8. https://doi.org/10.1109/IJCNN.2019.8852227

Bahdanau D, Cho K, Bengio Y, 2015. Neural machine translation by jointly learning to align and translate. 3rd Int Conf on Learning Representations.

Bernat E, Bunce S, Shevrin H, 2001. Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing. *Int J Psychophysiol*, 42(1):11-34. https://doi.org/10.1016/S0167-8760(01)00133-7

Cao JM, Li YY, Sun MC, et al., 2022. Do-Conv: depthwise over-parameterized convolutional layer. *IEEE Trans Image Process*, 31:3726-3736. https://doi.org/10.1109/TIP.2022.3175432

Chao H, Dong L, 2021. Emotion recognition using three-dimensional feature and convolutional neural network from multichannel EEG signals. *IEEE Sens J*, 21(2):2024-2034. https://doi.org/10.1109/JSEN.2020.3020828

Cheng J, Chen MY, Li C, et al., 2021. Emotion recognition from multi-channel EEG via deep forest. *IEEE J Biomed Health Inform*, 25(2):453-464. https://doi.org/10.1109/JBHI.2020.2995767

Goodfellow I, Pouget-Abadie J, Mirza M, et al., 2020. Generative adversarial networks. *Commun ACM*, 63(11):139-144. https://doi.org/10.1145/3422622

Guo JY, Cai Q, An JP, et al., 2022. A Transformer based neural network for emotion recognition and visualizations of crucial EEG channels. *Phys A Stat Mech Appl*, 603: 127700. https://doi.org/10.1016/j.physa.2022.127700

Hu JF, Min JL, 2018. Automated detection of driver fatigue based on EEG signals using gradient boosting decision tree model. *Cogn Neurodyn*, 12(4):431-440. https://doi.org/10.1007/s11571-018-9485-1

Jenke R, Peer A, Buss M, 2014. Feature extraction and selection for emotion recognition from EEG. *IEEE Trans Affect Comput*, 5(3):327-339. https://doi.org/10.1109/TAFFC.2014.2339834

Kingma DP, Welling M, 2014. Auto-encoding variational Bayes. 2nd Int Conf on Learning Representations.

Koelstra S, Muhl C, Soleymani M, et al., 2012. DEAP: a database for emotion analysis; using physiological signals. *IEEE Trans Affect Comput*, 3(1):18-31. https://doi.org/10.1109/T-AFFC.2011.15

Lan ZR, Sourina O, Wang LP, et al., 2016. Real-time EEG-based emotion monitoring using stable features. *Vis Comput*, 32(3):347-358. https://doi.org/10.1007/s00371-015-1183-y

Lew WCL, Wang D, Shylouskaya K, et al., 2020. EEG-based emotion recognition using spatial-temporal representation via Bi-GRU. 42nd Annual Int Conf of the IEEE Engineering in Medicine & Biology Society, p.116-119. https://doi.org/10.1109/EMBC44109.2020.9176682

Li C, Lin XJ, Liu Y, et al., 2022. EEG-based emotion recognition via efficient convolutional neural network and contrastive learning. *IEEE Sens J*, 22(20):19608-19619. https://doi.org/10.1109/JSEN.2022.3202209

Li JP, Zhang ZX, He HG, 2018. Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cogn Comput*, 10(2):368-380. https://doi.org/10.1007/s12559-017-9533-x

Li SJ, Li W, Xing ZJ, et al., 2022. A personality-guided affective brain–computer interface for implementation of emotional intelligence in machines. *Front Inform Technol Electron Eng*, 23(8):1158-1173. https://doi.org/10.1631/FITEE.2100489

Li X, Song DW, Zhang P, et al., 2016. Emotion recognition from multi-channel EEG data through convolutional recurrent neural network. IEEE Int Conf on Bioinformatics and Biomedicine, p.352-359. https://doi.org/10.1109/BIBM.2016.7822545

Li X, Zhang YZ, Tiwari P, et al., 2022. EEG based emotion recognition: a tutorial and review. *ACM Comput Surv*, 55(4): 79. https://doi.org/10.1145/3524499

Lin YP, Wang CH, Wu TL, et al., 2009. EEG-based emotion recognition in music listening: a comparison of schemes for multiclass support vector machine. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.489-492. https://doi.org/10.1109/ICASSP.2009.4959627

Liu YJ, Yu MJ, Zhao GZ, et al., 2018. Real-time movie-induced discrete emotion recognition from EEG signals. *IEEE Trans Affect Comput*, 9(4):550-562. https://doi.org/10.1109/TAFFC.2017.2660485

Liu YS, Sourina O, 2014. EEG-based subject-dependent emotion recognition algorithm using fractal dimension. IEEE Int Conf on Systems, Man, and Cybernetics, p.3166-3171. https://doi.org/10.1109/SMC.2014.6974415

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

Picard RW, 2000. Affective Computing. MIT Press, Cambridge, UK. https://doi.org/10.7551/mitpress/1140.001.0001

Salama ES, El-Khoribi RA, Shoman ME, et al., 2018. EEG-based emotion recognition using 3D convolutional neural networks. *Int J Adv Comput Sci Appl*, 9(8):329-337. https://doi.org/10.14569/IJACSA.2018.090843

Song TF, Zheng WM, Song P, et al., 2020. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans Affect Comput*, 11(3):532-541. https://doi.org/10.1109/TAFFC.2018.2817622

Sorkhabi MM, 2014. Emotion detection from EEG signals with continuous wavelet analyzing. *Am J Comput Res Repos*, 2(4):66-70. https://doi.org/10.12691/ajcrr-2-4-3

Stam CJ, 2005. Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. *Clin Neurophysiol*, 116(10):2266-2301. https://doi.org/10.1016/j.clinph.2005.06.011

Tang ZC, Li C, Wu JF, et al., 2019. Classification of EEG-based single-trial motor imagery tasks using a B-CSP method for BCI. *Front Inform Technol Electron Eng*, 20(8):1087-1098. https://doi.org/10.1631/FITEE.1800083

Tao W, Li C, Song RC, et al., 2023. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Trans Affect Comput*, 14(1):382-393. https://doi.org/10.1109/TAFFC.2020.3025777

Tripathi S, Acharya S, Sharma RD, et al., 2017. Using deep and convolutional neural networks for accurate emotion classification on DEAP data. Proc 31st AAAI Conf on Artificial Intelligence, p.4746-4752. https://doi.org/10.1609/aaai.v31i2.19105

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. 31st Int Conf on Neural Information Processing Systems, p.6000-6010.

Vijayan AE, Sen D, Sudheer AP, 2015. EEG-based emotion recognition using statistical measures and auto-regressive modeling. IEEE Int Conf on Computational Intelligence & Communication Technology, p.587-591. https://doi.org/10.1109/CICT.2015.24

Wang Q, Sourina O, Nguyen MK, 2011. Fractal dimension based neurofeedback in serious games. *Vis Comput*, 27(4): 299-309. https://doi.org/10.1007/s00371-011-0551-5

Wang XW, Nie D, Lu BL, 2014. Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129:94-106. https://doi.org/10.1016/j.neucom.2013.06.046

Wei C, Chen LL, Song ZZ, et al., 2020. EEG-based emotion recognition using simple recurrent units network and ensemble learning. *Biomed Signal Process Contr*, 58:101756. https://doi.org/10.1016/j.bspc.2019.101756

Yang BH, He LF, Lin L, et al., 2015. Fast removal of ocular artifacts from electroencephalogram signals using spatial constraint independent component analysis based recursive least squares in brain-computer interface. *Front Inform Technol Electron Eng*, 16(6):486-496. https://doi.org/10.1631/FITEE.1400299

Yang Y, Gao Q, Song XL, et al., 2021. Facial expression and EEG fusion for investigating continuous emotions of deaf subjects. *IEEE Sens J*, 21(15):16894-16903. https://doi.org/10.1109/JSEN.2021.3078087

Yang YL, Wu QF, Fu YZ, et al., 2018a. Continuous convolutional neural network with 3D input for EEG-based emotion recognition. 25th Int Conf on Neural Information Processing, p.433-443. https://doi.org/10.1007/978-3-030-04239-4_39

Yang YL, Wu QF, Qiu M, et al., 2018b. Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. Int Joint Conf on Neural Networks, p.1-7. https://doi.org/10.1109/IJCNN.2018.8489331

Yang YX, Gao ZK, Wang XM, et al., 2018. A recurrence quantification analysis-based channel-frequency convolutional neural network for emotion recognition from EEG. *Chaos*, 28(8):085724. https://doi.org/10.1063/1.5023857

Yin YQ, Zheng XW, Hu B, et al., 2021. EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. *Appl Soft Comput*, 100:106954. https://doi.org/10.1016/j.asoc.2020.106954

Zhang DL, Yao LN, Zhang X, et al., 2018. Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface. Proc 32nd AAAI Conf on Artificial Intelligence, p.1703-1710. https://doi.org/10.1609/aaai.v32i1.11496

Zhang QQ, Liu Y, 2018. Improving brain computer interface performance by data augmentation with conditional deep convolutional generative adversarial networks. https://arxiv.org/abs/1806.07108

Zhang T, Zheng WM, Cui Z, et al., 2019. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Trans Cybern*, 49(3):839-847. https://doi.org/10.1109/TCYB.2017.2788081

Zheng WL, Lu BL, 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans Auton Ment Dev*, 7(3):162-175. https://doi.org/10.1109/TAMD.2015.2431497

Zhong XY, Gu Y, Luo YT, et al., 2023. Bi-hemisphere asymmetric attention network: recognizing emotion from EEG signals based on the transformer. *Appl Intell*, 53(12): 15278-15294. https://doi.org/10.1007/s10489-022-04228-2