Self-supervised group meiosis contrastive learning for EEG-based emotion recognition

Haoning Kan · Jiale Yu · Jiajin Huang · Zihe Liu · Heqian Wang · Haiyan Zhou

Abstract

The progress of EEG-based emotion recognition has received widespread attention from the fields of human-machine interaction and cognitive science. However, recognizing emotions with limited labelled data is still challenging. To address this issue, this paper proposes a self-supervised group meiosis contrastive learning (SGMC) framework for EEG-based emotion recognition. First, to reduce the dependence of emotion labels, SGMC introduces a contrastive learning task according to the alignment of video clips based on the similar EEG response across subjects. Moreover, the model adopts a group projector to extract group-level representations from the group samples to further decrease the subject difference and random effects in EEG signals. Finally, a novel genetics-inspired data augmentation method, named meiosis is developed, which takes advantage of the alignment of video clips among a group of EEG samples to generate augmented groups by pairing, cross exchanging, and separating. The experiments show that SGMC exhibits competitive performance on the publicly available DEAP and SEED datasets. It is worth of noting that the SGMC shows a high ability to recognize emotion even when using limited labelled data. Moreover, the results of feature visualization suggest that the model might have learned video-level emotion-related feature representations to improve emotion recognition. The hyper-parametric analysis further shows the effect of the group size during emotion recognition. Finally, the comparisons of both the symmetric function and the ablation models and the analysis of computational efficiency are carried out to examine the rationality of the SGMC architecture. The code is provided publicly online.

Keywords EEG · Emotion recognition · Group-level representation · Contrastive-learning · Self-supervised learning · Data augmentation · Meiosis

1 Introduction

Emotion recognition is quite important in human-machine interaction and cognitive science. For example, users’ emotional states are evaluated while they are listening to music and watching movies, which aims to provide instruction to content creators [1, 2]. The emotion recognition method based on electroencephalography (EEG) signals has shown its advantages. When compared to conscious behaviour, such as facial expression and body language, EEG signals may reflect the internal thought or mind generally, since EEG signals are harder to cheat or disguise. When compared with other physiological signals, such as fMRI (functional magnetic resonance imaging), and ECG (electrocardiogram), EEG is more convenient for sampling and has a higher time resolution.

By carefully selecting the critical handcrafted features, there has been great progress in EEG-based emotion recognition [3–5], yet, these approaches rely too much on the researcher’s experiences with EEG signals and cognitive-related knowledge [6]. However, the development of deep learning methods and applications greatly promotes research on EEG-based emotion recognition, in which it is possible to learn high-level representations for emotion recognition [7–12].

Generally, a large number of manual labels are vital for training supervised models. To acquire the annotated dataset, subjects are usually required to assess their emotional situations during cognitive processing. However, obtaining emotion labels in this way is time-consuming and laborious. Additionally, it is easy to generate subjective bias, which may decrease the reliability of labels and limit their use [13,
To mitigate the issue of lacking labelled datasets, contrastive self-supervised learning (SSL) has emerged [15, 16] as a promising strategy. Contrastive SSL aims to train an encoder mapping the input data into an embedding space where the distances in the positive sample pairs are closer than those in the negative sample pairs. For example, based on the findings of consistent EEG responses across subjects from neuroscience [17, 18], Shen and colleagues adopted a SSL-based contrastive learning framework to maximize the similarity of EEG signals across subjects, which can learn the representation corresponding to the same emotional clip and improve the performance of emotion recognition [19]. However, random effects exist in the emotion-related EEG signals. For example, the possible distraction or fatigue situations of subjects during emotional tasks increases signal noise. Moreover, the responses of subjects cannot be exactly the same, and the difference across subjects is usually greater than the difference within a subject [20]. Individual difference would also increase the difficulty of maximizing the similarity across subjects in contrastive learning.

Hence, in this study, we propose a self-supervised group meiosis contrastive learning (SGMC) framework to further improve EEG-based emotion recognition under the SSL framework. Similar to the manipulation in Shen and colleagues’ work, SGMC applies a contrastive learning task based on the consistency of emotional stimuli, where the samples corresponding to the same video clips are positive pairs and the samples corresponding to different video clips are negative pairs. By considering the random effect in EEG signals and the difference across subjects, we design a group projector in SGMC to extract group-level representation. Research has found that increasing the group based sample size could improve the generality of the EEG-based generation model [21]. So to more reliably characterize the EEG signals, the group projector in SGMC collects the EEG samples from a group of subjects and extracts group-level emotional representations relating to the video clip.

Moreover, we propose a novel data augmentation method for contrastive learning. Usually applying the process of data augmentation to enhance contrastive learning is a basic paradigm, and the original input samples are often augmented by flipping, rotating, cropping and other operations. In this study, the proposed data augmentation method is inspired by the meiosis mechanism in genetics, and the EEG signals are augmented based on a group of samples. With the alignment of video clips, a group of EEG data are paired, cross-changed and separated, which can augment data without changing the original features. In this way, data augmentation enables contrastive learning to take advantage of the alignment of clips to construct positive and negative pairs. Hence, the SGMC can finally learn the critical representations related to the video clips and improve EEG-based emotion recognition performance.

Here we summarize the contributions of the proposed SGMC model in this paper as follows:

- To reduce the dependence of emotion labels, a contrastive learning task is designed based on the alignment of clips, which emphasizes the cross-subject consistency of EEG signals according to the same video clip.
- To decrease the subject difference and random effects in EEG signals, a group projector is designed to collect a group of samples corresponding to video clips and extract the group-level representations.
- To augment the group-based sample, a genetics-inspired data augmentation method, named meiosis, is developed. It utilizes the alignment of clips to augment EEG samples without changing the original features, and provides augmented group samples for contrastive learning.
- The SGMC achieves competitive emotion classification results on the publicly available DEAP and SEED datasets. In particular, it achieves high accuracy when fine-tuned with very limited labelled data, which even exceeds that of the supervised learning baseline model with a full training set.

2 Related work

2.1 EEG-based emotion recognition

As a nonstationary time series obtained from multiple channels, EEG signals contain abundant information in the temporal, spatial and frequency dimensions. In the research of EEG-based emotion recognition, one approach focuses on calculating and selecting the critical EEG handcrafted features related to emotion recognition, including the linear and nonlinear features in these dimensions [3–5, 22, 23].

Another approach is to learn deep representations in the temporal, spatial and frequency dimensions by directly feeding raw data into different models under the framework of deep learning. For example, recurrent neural networks (RNNs), long short term memory networks (LSTMs), and convolutional neural networks (CNNs) are quite popular. The RNNs and LSTMs focus on extracting temporal features [24], and can also extract spatial features by rearranging the input data. For example, RNN-based BiHDM adopts the inputs in each hemisphere from both horizontal and vertical streams [25]. CNNs are also used for extracting spatial features of EEG signals [26]. Similarly based on the rearrangement of inputs or the combination with RNNs, CNNs can also capture effective spatio-temporal features or temporal-frequency features [27–29]. For example, Yang et al. [28] designed a hybrid model combining CNN and RNN networks by utilizing the input combined with sparse
Self-supervised group meiosis contrastive learning for EEG-based emotion recognition

2.2 SSL

SSL aims to learn representation without relying on manual labels, and shows the capacity to achieve general and robust representation [35–41]. SSL is widely used in many fields. For example, in computer vision (CV), based on spatial features, Gidaris et al. [35] designed an SSL task to rotate the original images and required the model to predict the rotation angles. Based on the temporal features of the video, an SSL task [36] is designed to require the model to predict whether the two video frames are close in time. In natural language processing (NLP), there are several SSL tasks in word2vec [37], such as predicting headwords and adjacency words. In BERT [38], two SSL tasks are designed, including masked language prediction and next sentence prediction. In EEG signal processing, Zhang et al. [42] applied a generative adversarial network (GAN) to design the SSL task. Simulated signals are generated and augmented by masking original signals, and the model is required to distinguish real and simulated signals. The model shows state-of-the-art (SOTA) performance and alleviates the limitation of EEG data scarcity.

In recent years, contrastive-learning-based SSL has been introduced in EEG signal processing and exhibited great improvement during classification tasks. Contrastive learning defines two samples with internal relations as positive pairs; otherwise, they are negative pairs. Its loss function aims to maximize the similarities of representations between positive pairs and minimize the similarities between the negative pairs. To enhance the learning performance with limited labels, the positive and negative pairs can be constructed according to the time information. For examples, relative positioning (RP) requires the model to distinguish whether the two EEG signals are recorded closely in time or not, temporal shuffling (TS) requires the model to predict the time sequence of two samples [15], and TF-C(temporal-frequency consistency) embeds a time-based neighbourhood of an example close to its frequency-based neighbourhood [14]. With the physiologically and clinically meaningful feature representations by SSL, the models outperform the fully-supervised learning significantly when using less labelled data. As mentioned above, EEG pairs can be constructed based on the semantic perspective to improve the generalization ability across subjects [19]. Here, we adopt a similar contrastive learning task and the SGMC requires distinguishing whether two emotional EEG samples are triggered by the same video clip. However, our EEG samples are reconstructed based on group-level characteristics.

In addition, the data augmentation method is critical during contrastive learning. For example, Jiang et al. proposed [16] an augmentation-based SSL task, which requires the model to distinguish whether two augmented EEG signals come from the same original signal or not. Based on effective augmentation methods, such as time warping permutation and cropping and resizing, the generalization ability in sleep-stage recognition is increased, exceeding the performance of supervised learning. The data augmentation of meiosis is designed in our SGMC without changing the original EEG information, which is based on the group-based sample data.

3 Proposed method

3.1 Overall framework

As illustrated in Fig. 1, the proposed SGMC framework consists of a pre-training process based on contrastive learning and a fine-tuning process for emotion recognition. There are five components in the pre-training process: group sampler, meiosis data augmentation, base encoder, group projector, and contrastive loss function. The group sampler first generates a batch containing several groups of EEG signals for data augmentation. Meiosis augments each EEG group and generates two groups for constructing the positive and negative pairs. Next the base encoder extracts individual-level video clip-related representations from the signals. Then the group projector aggregates each group of representations to extract group-level video clip-related representations and maps them into another latent space for computing the similarity. In addition, the parameters of the base encoder and group projector are optimized by minimizing the contrastive loss. Finally, the pre-trained base encoder is fine-tuned with a classifier and performs emotion classification training during the fine-tuning process.

3.2 Group sampler

Generally, the feature representations extracted from individual EEG samples are difficult to use for contrastive learning. Therefore, we use the strategy of extracting from groups of EEG samples. To achieve this we construct the sampler to provide input for the batch.
In the processed dataset, video clips and subjects correspond to two axes of the dataset tensor. Each EEG sample is defined as $X^s_{v} \in \mathbb{R}^{M \times C}$, which corresponds to a 1-second signal recorded when Subject $s$ watched video $v$. $M$ is the length of the signal in time and $C$ is the electrode channel of signals. To obtain a batch, as illustrated in Fig. 2, the sampler first randomly samples $P$ video clips $v_1, v_2, \ldots, v_P$. To sample two equal sample groups and construct positive pairs for each clip, the sampler next randomly selects $2Q$ Subjects $s_1, s_2, \ldots, s_{2Q}$ for grouping. The sampler then extracts the EEG signals corresponding to selected subjects and video clips, where in $2PQ$ samples, a dataset of $D = \{X^s_{v_i} | i = 1, 2, \ldots, P; k = 1, 2, \ldots, 2Q\}$ is obtained, which are recorded by $2Q$ subjects watching $P$ video clips respectively. Furthermore, we note a group of samples $G_i = \{X^s_{v_1}, X^s_{v_2}, \ldots, X^s_{v_P}\}$ corresponding to video clip $v_i$. Among
of samples \( \{G_1, G_2, \ldots, G_P\} \) corresponding to different \( P \) video clips for pre-training.

### 3.3 Meiosis data augmentation

Meiosis aims to augment one group sample and generate two parts that preserve the same clip-related features for constructing the positive pairs.

We first mix signals from different subjects together for data augmentation. To preserve the original clip-related features, we select the signals corresponding to the same clip and use them for splitting and splicing. The function of crossover transformation is used to perform the processes, which are designed as follows:

We represent any EEG signal \( A \) as \( \{a_1, a_2, \ldots, a_M\} \), where \( a_i \) is the data at the \( i^{th} \) sampling point \( (i=1,2,\ldots,M) \). Similarly, we represent any other signal \( B \) as \( \{b_1, b_2, \ldots, b_M\} \). To preserve clip-related features, we utilize video clip alignment to align the consistent clips among samples, and then split and splice them into two samples along the same clip site. The data of the first \( N \) sampling points of the two samples \( A \) and \( B \) are exchanged to obtain \( \tilde{A} = \{b_1, b_2, \ldots, b_N, a_{N+1}, a_{N+2}, \ldots, a_M\} \) and \( \tilde{B} = \{a_1, a_2, \ldots, a_N, b_{N+1}, b_{N+2}, \ldots, b_M\} \), where \( N \) is given. Such transformations for any two EEG signals are encapsulated as the following function expression:

\[
\{\tilde{A}, \tilde{B}\} = T(A, B, N)
\]

Furthermore, to take advantage of group sampling, we randomly match pairs for crossover and separation. As illustrated in Fig. 3, the process of meiosis augmentation is designed as follows:

1) Individual pairing: For the original signals in the group \( G_i = \{X^v_{ki}|k = 1, 2, \ldots, Q\} \) (corresponding to the same video clip \( v_i \)), individual signals are randomly paired to form \( Q \) pairs \( \{X^{v_i}_{k1}, X^{v_i}_{k2}, \ldots, X^{v_i}_{kQ}\} \) for crossover.

2) Crossover: The split position \( N \) is randomly given and transformed to obtain pairs of signals \( \{\tilde{X}^{v_i}_{ki}, \tilde{X}^{v_i}_{k(i+Q)}\} \) for crossover.

3) Separation: The transformed signals are divided into two groups according to the split positions, and the paired signals are required to enter Group \( A \) and Group \( B \). Therefore, the two homologous groups of EEG \( \tilde{G}_i^A = \{\tilde{X}^{v_i}_{ki}|k = 1, 2, \ldots, Q\} \) and \( \tilde{G}_i^B = \{\tilde{X}^{v_i}_{k(i+Q)}|k = Q+1, Q+2, \ldots, 2Q\} \) are obtained and shared similar group-level clip-related features.

The data augmentation procedure for group samples is represented in the following formula:

\[
\{\tilde{G}_i^A, \tilde{G}_i^B\} = \text{meiosis}(G_i)
\]

Based on the meiosis, for the \( P \) groups of samples \( G \) in a batch, \( 2P \) groups of samples \( \tilde{G} \) can be obtained as follows:

\[
\tilde{G} = \{\tilde{G}_i|i = 1, 2, \ldots, P; t \in \{A, B\}\} = \text{meiosis}(G)
\]

The positive pairs are formed with \( \tilde{G}_i^A \) and \( \tilde{G}_i^B \), and the negative pairs are formed with the samples from any other \( 2(P - 1) \) groups of samples.
3.4 Base encoder

To extract group-level clip-related features for contrastive learning, we first design a base encoder to extract individual-level clip-related features from the EEG samples of individual subjects. Then we introduce the base encoder $f : \mathbb{R}^{M \times C} \rightarrow \mathbb{R}^D$ which maps EEG samples $X$ to its representation $h$ on a 512-dimensional feature space. Based on the existing model ResNet18-1D [30], the base encoder is designed as follows:

As illustrated in Fig. 4, it contains 17 convolutional layers (Conv) with a 1D kernel. The kernels of the first convolutional layer are paralleled to the time axis of the EEG signal tensor with a length of 9. Each residual block contains two convolutional layers with the same number and length of kernels. In each residual block, the kernels of the first layer are parallel to the time axis of the input EEG tensor, and the second layer is parallel to the channel axis. For the other eight residual blocks, the lengths of the kernels are 15, 15, 11, 11, 7, 7, 3, and 3 in descending order. In addition, the processes of max pooling with the 1D kernel (Maxpool), avg pooling with the 1D kernel (Avgpool), batch normalization (BN), and rectified linear unit (ReLU) layers are shown in the corresponding positions in Fig. 4.

Through the base encoder, for the augmented groups of samples, $\tilde{G}_t^i$, the individual-level clip-related representation set $\{h_1, h_2, ..., h_Q\}$ are obtained as follows:

$$H_t^i = f(\tilde{G}_t^i)$$  \hspace{1cm} (4)  

The set is used for further extracting group-level features. Individual representations are also used in extracting emotional features for classification.

3.5 Group projector

To eliminate random effects, such as fatigue and distraction states in subjects, the group projector is designed to extract group-level features from samples of multiple subjects. During this process, the representations from 1-second EEG signals are carefully projected into latent space, and then the similarity of video clips is calculated.

A group of samples is an unordered set of matrices. Most models are focused on regular input representations.
Fig. 4, the group projector comprises a base projector and features from group-based EEG signals. As illustrated in function to build a network and implement feature extraction. We introduce the base projector of individual-level representation are upgraded for extraction. A symmetric function of MaxPool1D.

To decrease individual-level feature loss, the dimensions of individual-level representation are upgraded for extraction. We introduce the base projector \( l : \mathbb{R}^D \rightarrow \mathbb{R}^H \) that adopts a multilayer perceptron (MLP) to project each individual-level representation \( h \) on a 4096-dimensional feature space. The base projector contains three fully-connected layers with 1024, 2048, and 4096 hidden units in ascending order and adopts ReLU as the activation function in the first two layers. The processes of BN and dropout with 0.5 are shown in the corresponding positions in Fig. 4.

To ensure an invariant output representing the group-based samples with any input permutations, 1-dimensional max pooling (MaxPool1D) is adopted to aggregate the information from each dimension-upgraded representation. As illustrated in Fig. 4, the 1D kernel of MaxPool1D is perpendicular to the dimension-upgraded representation vector. The scanning direction of the kernel is parallel to the upgraded representation vector with a stride of 1 and padding of 0. On each of the 4096 dimensions, MaxPool extracts the corresponding maximum values from the Q dimension-upgraded representations. Therefore, we obtain the group-level feature representations in the latent space.

The group projector is noted as \( g : \mathbb{R}^{Q \times D} \rightarrow \mathbb{R}^H \). The process of extracting group representations in latent space is obtained through \( g \) as follows:

\[
z^i_v = g(H^i_v) = \text{MaxPool1D}(l(h_1), l(h_2), ..., l(h_Q))
\]  

\( (5) \)

### 3.6 The contrastive loss

To measure the similarity between two groups of samples, we calculate the cosine similarity of their group-level representation vectors. The input of group-based samples \( \{G^i_t | i = 1, 2, ..., P; t \in \{A, B\} \} \) are extracted to obtain group-level feature representations \( \{z^i_t | i = 1, 2, ..., P; t \in \{A, B\} \} \) via the base encoder and group projector. Then, the similarity of the samples in the two augmented Groups \( G^i \) and \( G^j \) are calculated according to \( z^A_i \) and \( z^B_j \):

\[
s(z^A_i, z^B_j) = \frac{z^A_i \cdot z^B_j}{\|z^A_i\| \|z^B_j\|}, \quad s(z^A_i, z^B_j) \in [0, 1]
\]  

\( (6) \)

The contrastive loss is designed to maximize the similarity of two group-level representations in a positive pair and minimize the similarity in the negative pairs. Similar to the SimCLR framework \( (44) \), we adopt the normalized temperature-scaled cross-entropy to define the loss function as follows:

\[
\ell^A_i = -\log \frac{\exp(s(z^A_i, z^B_j)/\tau)}{\sum_{j=1,j \neq i}^P \exp(s(z^A_i, z^B_j)/\tau)} + \frac{\sum_{j=1}^P \exp(s(z^A_i, z^B_j)/\tau)}{\sum_{j=1,j \neq i}^P \exp(s(z^A_i, z^B_j)/\tau)}
\]  

\( (7) \)

where \( 1_{[j \neq i]} \in \{0, 1\} \) is an indicator function equal to 1 if \( j \neq i \). \( \tau \) is the temperature parameter of softmax. The smaller the loss function is, the larger the similarity between \( z^A_i \) and \( z^B_j \) is, and accordingly the smaller the similarity between \( z^A_i \) and other group representations is.

Finally, we define the total loss for an iteration as the average of all contrastive losses for backpropagation as follows:

\[
L = \frac{1}{2P} \sum_{i=1}^P (\ell^A_i + \ell^B_i)
\]  

\( (8) \)

### 3.7 Classifier

In the next fine-tuning task, we use the classifier to extract emotional features and predict emotion labels based on the representations extracted by the base encoder. As illustrated in Fig. 4, the classifier mainly contains three fully-connected layers with 512, 256, and 128 hidden units in descending order. Batch normalization ReLU and dropout with 0.5 are shown in the corresponding positions in Fig. 4.

### 3.8 Pre-training process

The SGMC pre-training is performed based on the constructed group sampler, data augmentation, base encoder, group projector, and loss function.

During the pre-training, we first set a number of epochs \( T_1 \), and then iterate the epochs. In each epoch, we continue to sample \( P \) video clips per iteration until all video clips are enumerated. In each iteration, the sampler extracts \( 2PQ \) EEG samples \( D = \{X^i_k | i = 1, 2, ..., P; k = 1, 2, ..., 2Q\} \) and packs them into \( P \) Groups \( G = \{G^i_t | i = 1, 2, ..., P\} \). Next, for meiosis data augmentation, to avoid model cheating by recognizing the split position, we randomly generate a fixed split position \( N \), and send it to each meiosis in this iteration \( (1 < N < M - 1) \). \( 2P \) augmented group samples...
\( \mathcal{G} = \{ \mathcal{G}_t | t = 1, 2, \ldots, P; t \in \{ A, B \} \} \) can be obtained by Formula (3). Furthermore, we

\begin{algorithm}
\textbf{Algorithm 1} The pre-training process.
\begin{algorithmic}[1]
\State \textbf{Input:} Number of video clips \( P \) per batch, number of subjects \( Q \) per group. Initilized base encoder \( f \) and group projector \( g \).
\For {epoch = 1 to \( T_1 \)}
\Repeat
\State Sample \( P \) video clips \( \{v_i | i = 1, 2, \ldots, P \} \).
\State Randomly select \( 2Q \) subjects \( \{ s_k | k = 1, 2, \ldots, 2Q \} \).
\State Sampler pack batch \( \mathcal{G} = \{ \mathcal{G}_t | t = 1, 2, \ldots, P \} \) from \( \mathcal{D} = \{ X^n_t | i = 1, 2, \ldots, P; k = 1, 2, \ldots, 2Q \} \).
\State Randomly generate a split position \( c \).
\State Obtain \( \mathcal{G} = \{ \mathcal{G}_t | i = 1, 2, \ldots, P; t \in \{ A, B \} \} \) from \( \mathcal{G} \) through meiosis with \( N \) by (1)-(3).
\State Obtain \( Z = \{ x^n_t | i = 1, 2, \ldots, P; t = \in \{ A, B \} \} \) from \( \mathcal{G} \) through \( f \) and \( g \) by (4) and (5).
\State Calculate loss \( L \) by (6)-(8).
\State Abate loss \( L \) through optimizer updating parameters of \( f \) and \( g \).
\Until {all video clips are enumerated.}
\EndFor
\State \textbf{Output:} base encoder \( f \), throw away group projector \( g \).
\end{algorithmic}
\end{algorithm}

extract group-level features and project them to latent space to obtain group representations by Formulas (4) and (5). In addition, we calculate the loss \( L \) by Formulas (6)-(8). Finally, to reduce the loss \( L \), we calculate the gradient through the backpropagation method to update the optimizer parameters \( f \) and \( g \). Detailed procedures are summarized in Algorithm 1.

3.9 Fine-tuning process

Based on the feature representations learned during the pre-training process, we further fine-tune the model with labelled samples to improve the emotional classification performance. As illustrated in Fig. 1, emotion classification with supervised training is performed on the model consisting of an initialized classifier and the SGMC pre-trained base encoder.

\begin{algorithm}
\textbf{Algorithm 2} The fine-tuning process.
\begin{algorithmic}[1]
\State \textbf{Input:} Training data \( \{ X \} \), Training labels \( \{ y \} \), Pre-trained base encoder \( f \), initialized classifier \( k \).
\For {epoch = 1 to \( T_2 \)}
\Repeat
\State Sample one batchsize of training samples \( X_B, y_B \) from \( \{ X \} \), \( \{ y \} \).
\State Calculate individual feature vectors \( h_B = f(X_B) \).
\State Obtain predicted labels \( y_B^{pre} = k(h_B) \).
\State \( L^{pre} = CE(y_B^{pre}, y_B) \), where \( CE(\cdot) \) means calculating the cross-entropy loss.
\State Abate loss \( L^{pre} \) through optimizer updating parameters of \( f \) and \( k \).
\Until {all samples in \( \{ X \} \) have been sampled.}
\EndFor
\State \textbf{Output:} base encoder \( f \) and classifier \( k \).
\end{algorithmic}
\end{algorithm}

We first set a number of epochs \( T_2 \) and denote the training data as \( X \) and their labels as \( y \). We denote the classifier as \( k(\cdot) \) and the label \( y \) is a categorical variable. For example, if there are four categories of emotion, the values of \( y \) are: 0, 1, 2 or 3. We need to predict the emotion category \( y \) for each sample \( X \in \mathbb{R}^{M \times C} \). The pre-trained base encoder \( f \) extracts the representation from the original EEG signal \( X \) for classifier \( k(\cdot) \), which further extracts the predictive features to obtain prediction categories \( y^{pre} = k(f(X)) \). We apply the cross-entropy function to define the loss function for the emotion classification task and apply an optimizer to minimize the loss function by optimizing the parameters of the model. Finally, when the loss function converges, a predictive EEG-based emotion recognition model is obtained. Detailed procedures are summarized in Algorithms 2 and 3.

\begin{algorithm}
\textbf{Algorithm 3} The prediction process.
\begin{algorithmic}[1]
\State \textbf{Input:} Data \( \{ X \} \), Fine-tuned base encoder \( f \) and classifier \( k \).
\State \textbf{Output:} predicted labels \( y^{pre} = k(f(X)) \).
\end{algorithmic}
\end{algorithm}

4 Experiments

In this section, we introduce the implementation details on the publicly available DEAP and SEED datasets. Furthermore, we carry out several experiments to explore the performance and mechanism of the proposed SGMC. First, we compare the performance of SGMC with other competitive methods in emotion recognition. Then we further evaluate its performance on limited labelled sample learning.

4.1 Implementation details

We first introduce the DEAP and SEED datasets and data processing. Then the basic hyperparameters utilized in the model are introduced later.

4.1.1 Datasets

DEAP: The widely-used DEAP dataset includes 32-channel EEG signals and 8-channel peripheral physiological signals recorded from 32 subjects who are watching emotional videos. There are a total of 40 video trials. Each EEG trial includes a 3-second resting signal and a 60-second emotional video clip-related signal. The recorded EEG signals are downsampled to a 128 Hz sampling rate and processed with a bandpass frequency filter of 4-45 Hz by the provider. Subjects are asked to rate their emotional levels of arousal, valence, liking, and dominance from 1 to 9 for each video after watching. We utilize the EEG signals and rating values of arousal and valence for the task of emotion recognition.
When the rating values are greater than 5.0, the corresponding EEG signals are labelled as high arousal or high valence. Otherwise, the signals are labelled as low arousal or low valence. Based on the labels, we perform binary classification tasks on both valence and arousal dimensions and a four-category classification task combining the two dimensions.

SEED: The SEED dataset is also widely used in emotion recognition. The dataset records the EEG signals from 15 subjects watching 15 movie videos, and the emotion labels are positive, neutral, and negative. Each video trial is approximately 4 minutes long. Each subject repeated three sessions of watching the same videos, with an interval of more than one week. The EEG signals are recorded via 62 electrodes with a sampling rate of 1000 Hz, resampled to 200 Hz and filtered in the range of 0-75 Hz by the provider.

4.1.2 Data processing

On DEAP, for each trial, we use a 1s sliding window to divide the 63s signal into 63 nonoverlapping segments. Then the 1s emotion-related signals minus the average of the first 3s resting signal are used as input data [28]. There are a total of 2400 segments of signals corresponding to 2400 video clips (40 videos with 60 1s clips) obtained from each subject. Then we divide EEG signals corresponding to the same video clips into the same set.

On SEED, we first perform L2 normalization in each channel. Then, similar to data processing on DEAP, for each trial, we divide the EEG signals into 1s segments without overlapping (discard the remaining segments shorter than 1s). Therefore, for each subject, there are 10182 segments of EEG signals from the 3 sessions, which correspond to 3394 1s video clips.

4.1.3 Basic configuration

In our strategy of evaluating the pre-training framework, two steps obtain its final emotion recognition results. We first save pre-trained models with different epochs. Next, we select the model with the highest average accuracy on emotion recognition obtained from five fine-tuning iterations. Such average accuracy is evaluated as a result of the pre-training framework.

We use PyTorch [45] to implement our experiments based on the NVIDIA RTX3060 GPU. The Adam optimizer [46] is used to minimize the loss functions for both the pre-training and fine-tuning processes. We represent \( lr \) as the learning rate of the optimizer. In the pre-training process and fine-tuning process, the number of epochs, batch size, temperature parameter \( \tau \), learning rate \( lr \), number of video clips per iteration \( P \), number of samples per Group \( Q \), and size of the dataset tensor are applied with different values. All the hyperparameters used in the processes on DEAP and SEED are listed in Table 1.

4.2 Emotion classification performance

4.2.1 General performance on DEAP

On the DEAP dataset, we first compare the SGMC with four SOTA methods in the two emotion dimensions of valence and arousal: CNN-LSTM [28], a hybrid network combined with convolutional networks and long short-term memory networks; MMResLSTM [47], a residual long short-term memory network utilizing multimodal data; CDCN [48], a channel-fused dense convolutional network; and ACRNN [49], a hybrid network of convolutional neural networks and

| Hyper Parameters | Pre-training | Fine-tuning | Pre-training | Fine-tuning |
|------------------|--------------|-------------|--------------|-------------|
| **Epoch**        | 2800         | 60          | 3288         | 70          |
| **batchsize**    | 32           | 2048        | 64           | 256         |
| **lr**           | \(10^{-4}\)  | \(10^{-3}\) | \(10^{-3}\)  | \(10^{-3}\) |
| \(\tau\)         | \(10^{-1}\)  | -            | \(10^{-1}\)  | -            |
| \(P\)            | 8            | -            | 16           | -            |
| \(Q\)            | 2            | -            | 2            | -            |
| \(Shape_{tr}\)   | (1680, 32, 1, 32, 128) | (53760, 1, 32, 128) | (2374, 45, 1, 62, 200) | (106380, 1, 62, 200) |
| \(Shape_{te}\)   | (360, 32, 1, 32, 128) | (11520, 1, 32, 128) | (510, 45, 1, 62, 200) | (22950, 1, 62, 200) |
| \(Shape_{val}\)  | (360, 32, 1, 32, 128) | (11520, 1, 32, 128) | (510, 45, 1, 62, 200) | (22950, 1, 62, 200) |

\(^{a}\) The **Epoch** represents an appropriate number of the pre-training or fine-tuning epochs for achieving a satisfying performance during emotion recognition and the **batchsize** represents the number of samples in a batch in the experiment on the general performance of emotion classification

\(^{b}\) The items of \(Shape_{tr}\), \(Shape_{te}\), \(Shape_{val}\) represent the size of tensors of training, test and the validation datasets for pre-training and fine-tuning respectively in the experiment on the performance with limited labelled sample learning, analysis of model performance improvement and analysis of architecture design.
Table 2 Performances on DEAP

| Methods                  | Valence | Arousal | Four          |
|-------------------------|---------|---------|---------------|
| CNN-LSTM (2020) [28]    | 90.82   | 86.13   | -             |
| CDCN (2020) [48]        | 92.24   | 92.92   | -             |
| MMResLSTM (2019) [47]   | 92.87   | 92.30   | -             |
| ARCNN (2019) [49]       | 93.72   | 93.38   | -             |
| MCLFS-GAN (2020) [50]   | -       | -       | 81.32         |
| GANSER (2022) [42]      | 93.52   | 94.21   | 89.74         |
| Proposed (Fully-supervised) | 92.25 | 92.87  | 88.88         |
| Proposed (Fine-tuning)  | 95.31   | 95.79   | 93.42         |

* Average accuracies (%) with SOTA methods and proposed frameworks on the DEAP dataset for valence classification, arousal classification and four classification.

Recurrent networks with a channelwise attention mechanism. Five-fold cross-validation is used to evaluate the performance of the proposed and comparative models. And the average performance of the 5-fold validation process is taken as the final experimental result. As shown in Table 2, the accuracy of the proposed SGMC is 1.59% higher than the second highest accuracy of the model in the valence dimension, and 1.58% higher in the arousal dimension. The comparison results demonstrate the effectiveness of the SGMC on EEG-based emotion recognition.

To verify the effectiveness of the proposed framework in the data augmentation and self-supervised learning fields, we further compare the performance of four-category classification in SGMC with MCLFS-GAN [50], a GAN-based data augmentation method, and GANSER [42], a self-supervised GAN-based data augmentation framework. In Table 2, the results show that the proposed SGMC outperforms the other two models by 12.10% and 3.68% on four-category classification respectively. As illustrated in Fig. 5, SGMC achieves high performance in each category, especially in low arousal and high valence.

Finally, we compare the proposed SGMC with our own fully-supervised learning baseline using the same network model without pre-training. In valence, arousal, and four-category dimensions, the accuracies of the SGMC exceed the fully-supervised learning baseline model by 3.06%, 2.92% and 4.54% respectively, suggesting the improvement of pre-training.

4.2.2 General performance on SEED

Similar to DEAP, we first compare our proposed SGMC with four supervised SOTA studies: GRSLR [51], adopting a graph regularized sparse linear regression model; BiHDM [25], utilizing two independent recognition networks for the left and right hemispheres of the brain; DGCNN [52], adopting a dynamic graph convolutional neural network; and ResNet18 [30], a 1D CNN-based residual neural network. Similarly, 5-fold cross-validation is used to evaluate the performance of the proposed and comparative models. And the average performance of the 5-fold validation process is taken as the final experimental results. As illustrated in Table 3, the accuracy of the proposed SGMC is 1.53% higher than the second-highest accuracy of the model. SGMC outperforms the four SOTA studies, and reflects its good emotion recognition performance on SEED. The results of accuracy in the classification task on positive, neutral, and negative emotions are illustrated in Fig. 5b. The SGMC achieves good accuracies in three categories, especially in positive emotion classification.

Similarly, we compare the SGMC with our fully-supervised learning baseline using the same model. When using 100% percentage of the labelled data, SGMC significantly exceeds our fully-supervised learning baseline by 3.32%, which suggests that SGMC pre-training contributes an obvious improvement. Moreover, when 1%, 10%, and 50% percentage of the labelled data are used, SGMC exceeds the corresponding fully-supervised learning baseline by 49.99%, 31.79%, and 7.00% respectively. The results suggest that proposed SGMC makes up for the lack of manual labels by utilizing large unlabelled data.
Table 3 Performances on SEED

| Methods                                      | Percentage of labels | 1%   | 10%  | 50%  | 100% |
|----------------------------------------------|----------------------|------|------|------|------|
| GRSLR (2018) [51]                           | -                    | -    | -    | 87.39|      |
| DGCNNN (2018) [52]                          | -                    | -    | -    | 90.40|      |
| BiHDM (2019) [25]                           | -                    | -    | -    | 93.12|      |
| ResNet18 1D kernel (2021) [30]              | -                    | -    | -    | 93.43|      |
| Proposed (Fully-supervised)                 | 42.59                | 62.49| 87.63| 91.64|      |
| Proposed (Fine-tuning)                      | 92.58                | 94.28| 94.63| 94.96|      |

* Average accuracy(%) with SOTA methods and proposed frameworks on the SEED dataset for positive, neutral and negative three-classification. Percentage of labels represents the ratio of the labelled data used in the fine-tuning or fully-supervised learning process.

4.2.3 Performance on limited labelled sample learning

To further evaluate the performance on limited labelled sample learning, we carry out emotion classification on both DEAP and SEED with a gradually increasing number of labelled data per category. To reduce the workload caused by the thousands of epochs during the pre-training, we use the validation dataset to adjust the hyperparameters of the framework and use the test dataset to evaluate it. The data samples are randomly divided into a training set, validation set and testing set at a ratio of 70:15:15. Moreover, to speed up the sampling, we shape the dataset tensor into different shapes in the pre-training and fine-tuning process. In the pre-training process, to facilitate the group sampler extracting groups of samples based on video clips, we set the five axes of the dataset tensor to correspond to the video clip, subject, 1, channel, and sampling point. In the fine-tuning process, the first two axes of the video clip and subject of the dataset are reshaped into a sample axis. Then, each axis of the reshaped dataset corresponds to the sample, 1, channel, and sampling point in turn. The tensor shapes of the training set, testing set, and validation set are represented as Shape_tr, Shape_te, and Shape_val and are listed in Table 1.

Similarly, we compare the results with the performance between the SGMC model pre-trained with the full training set and the fully-supervised learning model without pre-training. On DEAP, the models conduct a four-category classification combining arousal and valence information, and on SEED, the models conduct a three-category classification. As illustrated in Fig. 6, the accuracies of SGMC fine-tuning are superior to those of the fully-supervised learning baseline model on both DEAP and SEED. On DEAP, after the number of labelled samples per category is over 10, the accuracy gap between the SGMC and the fully-supervised learning method increases. When fine-tuned with 5000 labelled samples per category (37.2% of the full training set), the SGMC reaches a high accuracy of 87.51%, which is close to the accuracy of 87.68% in the fully-supervised learning training with the full training set. On SEED, when fine-tuned with only one labelled sample per category (0.00278% of the training set), the SGMC reaches an accuracy of 59.42%. When fine-tuned with 50 samples
per category (0.14% of the training set), the accuracy of 91.01% outperforms the fully-supervised learning baseline model with 100% labelled data. Furthermore, when the number of labelled samples per category is over 50, the SGMC performance line seems to converge to the limit and no longer increases. This result indicates that SGMC reduces the demand for manual labels. In addition, compared to the performance on DEAP, there is a superior classification ability on SEED with limited labelled samples. We speculate that this might be related to the larger sample size since the sample size in SEED is almost twice that in DEAP.

5 Discussion

In this part, we perform an analysis to discuss the reason for the performance improvement and the rationality of the proposed SGMC. First, in the analysis of model performance improvement, we analyse the model effectiveness by visualizing the feature representation learned by the SGMC. Additionally, we explore the meaningful law of the framework by evaluating the different combinations of hyperparameters. Second, to verify the rationality of the architecture design, both the comparisons of the symmetric function and the ablation models are conducted. Furthermore, the computational efficiency of proposed SGMC is discussed.

5.1 Analysis of model efficiency improvement

5.1.1 Representation visualization

To explore why the SGMC exhibits high emotion recognition ability, we visualize the learned feature representations of the SGMC fine-tuned model and the only fully-supervised learning baseline model on SEED through t-SNE [53].

As illustrated at the top of Fig. 7, the representations corresponding to three categories of emotion in t-SNE embedded space are marked with three colours. The distribution between each pair of emotions extracted by the SGMC model (right) is much more distinguishable than that based on the fully-supervised learning model (left). Moreover, there are some separated clusters in the representations with SGMC. When we mark the representations corresponding to the 15 movie videos with 15 different colours, it is obvious that the clusters in the representation of SGMC correspond to the videos (right in the bottom of Fig. 7). There are no such video-based clusters with the fully-supervised learning model.
learning model (left in the bottom of Fig. 7). These distribution patterns indicate that the SGMC model might have learned the video-related feature representations, which is beneficial to extracting emotional features.

5.1.2 Effect of hyper parameters

In this section, we further explore the effect of the number of samples per group ($Q$) and the number of selected video clips per iteration ($P$) by evaluating various combinations of hyperparameters. In our experimental strategy, for a given $Q$, we vary the $P$ values to 2, 4, 8, 16, 32 and 64 and select the $P$ value with the best accuracy in emotion recognition for the given $Q$. The results of the $Q$ values on emotion recognition are illustrated in Fig. 8. Similarly, four-category classification and three-category classification are carried out on DEAP and SEED respectively. In addition, the corresponding hyper parameters, including the $Q$, $P$, numbers of epochs of pre-training, and accuracies of pre-training are reported in Table 4.

First when the group size of $Q = 2$ is selected, the emotion recognition accuracy outperforms the condition with individual samples ($Q = 1$) on both datasets. On DEAP, when $Q = 2$ and $P = 8$, the SGMC reaches the best performance, and on SEED, when $Q = 2$ and $P = 16$, the performance is the best. These results show the advantages of group-level samples over individual-level samples with contrastive learning. Furthermore, there is a trend that the greater the group size $Q$ (when $P$ is constant), the greater the pre-training accuracy. The larger group sample contains more comprehensive group-level clip-related features, which might weaken the interference of random distractions, fatigue, and individual differences. In addition, our experimental results indicate that good accuracy in pre-training is not always beneficial to emotion recognition. Sometimes, the performance is even worse than that using individual-level samples ($Q = 1$). The possible reason is that the model might be focused on the aggregation of group-level features, which leads to the base encoder ignoring the related individual-level features. Therefore, it is critical to explore an appropriate group size for constructing group-based samples.

Second, there seems to be a different law in the DEAP and SEED datasets. When given a larger $Q$, the best $P$ value on DEAP tends to be smaller, and on SEED the best $P$ value tends to be larger. On DEAP, the emotional labels are determined by the rating of the subjects. On SEED, the labels are determined by the experimental designer, which is less related to the individual-based emotional response. When $P$ is larger, contrastive learning is harder since there are more pairs to match. This difficulty might encourage the model with a larger group size to focus on extracting group-level features and ignore the irrelevant individual-level features, which improves the performance on SEED.

![Fig. 8](image-url) The group size effect on the accuracy of DEAP (left) and SEED (right). The x-axis represents the number of samples ($Q$). The four lines with colors show the various percentages of labelled samples used for fine-tuning in the training set.

| Table 4 | Illustration of the appropriate combination of hyper parameters of $Q$ and $P$ in the hyper parameter analysis on DEAP and SEED |
|---------|--|---|---|---|---|
| DEAP    | $Q$ | $P$ | $Epoch_{pre}$ | $acc_{pre}$ | SEED    |
|         |     |     |               |           |         |
| 1       | 16  | 440 | 70.56         | 1992      | 71.58   |
| 2       | 8   | 2800| 91.11         | 3288      | 80.68   |
| 3       | 8   | 3600| 87.50         | 1296      | 66.24   |
| 4       | 4   | 800 | 93.06         | 744       | 69.82   |
| 8       | 4   | 475 | 96.94         | 548       | 72.45   |
| 16      | 4   | 450 | 97.08         |           |         |

$*Epoch_{pre}$ represents the appropriate number of epochs of pre-training, $acc_{pre}$ represents the probability that the similarity of positive pair is greater than that of all negative pairs, $Q$ represents number of samples per group, $P$ represents number of sampled video clips per iteration.
5.2 Architecture design analysis

In this section, to evaluate the architecture design, first the effects with different symmetric functions are compared. Then the ablation experiment is conducted by considering the important components. Finally, the computational efficiency is analyzed and discussed.

5.2.1 Comparison of symmetric functions

The proposed SGMC uses the symmetric function of MaxPool1D to construct the group projector. To verify its rationality, we compare MaxPool1D with the commonly used function of AvgPool1D. We also compare it with a MinPool1D function, which has an opposite feature selection ability by taking the minimum value from the upgraded representations. As illustrated in Fig. 9 and Table 5, the performance with MaxPool1D is better than the other two, suggesting that the function of MaxPool1D has the advantage of extracting clip-related features.

5.2.2 Ablation Experiment

To further investigate the rationality of the SGMC architecture, we conduct an ablation experiment based on three components: group-based sampling, meiosis data augmentation, and video clip consistency-based pair construction. We operate the new versions by removing one or two components, and the evaluation strategy is consistent with the basic configuration. In general, when the application of group sampling is ablated, we use individual-based samples for contrastive learning (just let $Q = 1$). When the meiosis data augmentation process is ablated, we skip the crossover process and go directly into the separation process after completing the individual-level pairing. After removing the video clip consistency-based pair construction, we change the method of constructing the positive pairs according to the consistent clips. The group sampler extracts a group of samples corresponding to any $2Q$ clip to generate the positive pairs via meiosis. The sampler is required to randomly sample EEG signals with any clip to form the sample groups.

The results of the four-category classification on DEAP and three-category classification on SEED are reported in Fig. 10. The details of ablation and the number of epochs of pre-training are reported in Table 6.

With NoGroup, which removes the group-based sampling, the performances on DEAP and SEED both declined by approximately 1.5%, suggesting the importance of group sampling in the contrastive learning of the SGMC framework. The accuracy of the NoAug version by removing meiosis augmentation decreases 3% on DEAP and 1.2% on SEED, suggesting the critical role of meiosis data augmentation in the SGMC in improving emotion recognition. To verify the effectiveness of meiosis utilizing the video clip alignment in the group sample, we design a comparative version with MixupAug. The data augmentation method of mixup [54] is used to generate mixed samples by weighted adding the two EEG signal matrices, which ignores clip alignment. We construct the MixupAug version by substituting the crossover process in meiosis with the mixup method. The results show that the meiosis-based SGMC exceeds the MixupAug-based model by 2.3% on DEAP and 2.6% on SEED, which suggests the effectiveness of meiosis augmentation by taking advantage of clip alignment. To verify the importance of constructing positive and negative pairs based on the consistent clips, we design a version of NoCon by

| Symmetric function | $\text{Epoch}_{\text{pre}}$ |
|--------------------|------------------------|
|                    | DEAP       | SEED       |
| MinPool1D          | 2440       | 1480       |
| AvePool1D          | 1720       | 2472       |
| MaxPool1D          | 2880       | 3288       |
5.2.3 Computational Efficiency Analysis of SGMC

To further analyze the computational efficiency of SGMC, we first list the calculations of time complexity in the proposed SGMC and the fully-supervised learning baseline model in Table 7. The $\theta_B$, $\theta_G$, $\theta_C$ represents the numbers of parameters in base encoder, group projector, classifier, respectively. The $n_v$ is the total number of video clips in training set for pre-training in SGMC and $n$ is the total number of samples in training set for fine-tuning in SGMC or in the fully-supervised learning model. And the $P$ is the number of video clips per batch during pre-training and the $B$ is the number of samples in a batch during fine-tuning or in the baseline model. Furthermore, the $T_1$ represents the number of epochs in pre-training and $T_2$ represents the number of epochs in fine-tuning or in the baseline model, respectively. During the pre-training stage, there are a total of $n_v$ video clips and all the video clips are required to be sampled in an epoch, so the time complexity is $n_vP$ times of a batch complexity for an epoch and $T_1n_vP$ times of batch complexity for the entire pre-training process. Similarly, the time complexity is $nB$ times of a batch complexity for an epoch and $T_2nB$ times of a batch complexity for the entire fine-tuning process in SGMC or in the fully-supervised learning baseline model, since their sampling methods are based on batch. Therefore, by comparing to the fully-supervised learning baseline model, the time complexity in the proposed SGMC is mainly from the additional pre-training process.

Table 6  The design of the five new versions and the complete SGMC, and their appropriate numbers of epochs for pre-training with each version on DEAP and SEED

| Methods     | Group | Augmentation | Consistency | $\text{Epoch}_{\text{pre}}$ |
|-------------|-------|--------------|-------------|-----------------------------|
|             |       |              |             |DEAP | SEED |
| NoGroup     | ✗     | meiosis      | ✓           | 440  | 1848 |
| NoAug       | ✓     | ✗            | ✓           | 800  | 2280 |
| MixupAug    | ✓     | mixup        | ✓           | 275  | 1304 |
| NoCon       | ✓     | meiosis      | ✗           | 60   | 2752* |
| OnlyCon     | ✗     | ✗            | ✓           | 1740 | 2368 |
| SGMC        | ✓     | meiosis      | ✓           | 2800 | 3288 |

* NoCon leads to worse performance of emotion recognition than fully-supervised on the SEED dataset, so we adopt the result obtained when the loss function of pre-training converges
b $\text{Epoch}_{\text{pre}}$ represents the appropriate number of epochs of pre-training
c mixup represents adopting a method of mixup to substitute crossover in meiosis data augmentation
Then, we focus on analyzing the computational efficiency of pre-training process in SGMC. According to the calculations of time complexity in Table 7, the key factor is the number of epoch and the time cost in an epoch. As listed in Table 4, the required numbers of epoch in pre-training process to ensure the highest performance in fine-tuning process are varied according to the combination of hyperparameters. We take the conditions with $Q = 2$ and $P = 8$ in DEAP and with $Q = 2$ and $P = 16$ in SEED as examples. As shown in Fig. 11, the model performance increases quickly in the beginning and then stops obvious increasing approximately after 500 epochs of training on DEAP and 1500 epochs on SEED. Moreover, as listed the time cost in Table 8, the proposed SGMC spends a little more time of 11.762s for DEAP and 30.363s for SEED for an epoch in the additional pre-training process. However, as discussed above, the pre-training in SGMC contributes greatly to the improvement of emotion classification performance.

### Table 7 Calculations of complexity in the proposed SGMC and fully-supervised learning baseline model

|             | Proposed (Pre-training) | Proposed (Fine-tuning) | Proposed (Fully-supervised) |
|-------------|-------------------------|------------------------|----------------------------|
| Batch       | $O(\theta_B + \theta_G)$ | $O(\theta_B + \theta_C)$ | $O(\theta_B + \theta_C)$ |
| Epoch       | $O\left(\frac{\tau_n}{k}(\theta_B + \theta_G)\right)$ | $O\left(\frac{\tau_n}{k}(\theta_B + \theta_C)\right)$ | $O\left(\frac{\tau_n}{k}(\theta_B + \theta_C)\right)$ |
| Total       | $O\left(\frac{T_1}{k}\tau_n(\theta_B + \theta_G)\right)$ | $O\left(\frac{T_2}{k}\tau_n(\theta_B + \theta_C)\right)$ | $O\left(\frac{T_2}{k}\tau_n(\theta_B + \theta_C)\right)$ |

### 6 Conclusion and future work

In this work, we propose SGMC framework to improve the EEG-based emotion recognition. In the proposed framework, meiosis data augmentation is introduced to augment EEG groups of samples without changing video clip features. A base encoder and a group projector are designed in the model to extract group-level feature representations. With the consistency of video clips, contrastive learning is designed to learn clip-related feature representations.

The proposed framework achieves superior emotion recognition results on both the DEAP and SEED datasets. Compared to the fully-supervised learning baseline, the SGMC improves emotion recognition significantly, especially when there are limited labelled data. In addition, the results of feature visualization suggest that the model might have learned the video-level emotional representations to improve model performance. The hyper parametric analysis further demonstrates the role of group-level samples during emotion recognition. Finally, the rationality of the framework design including the selection of symmetric functions, the construction of positive-negative pairs, and meiosis data augmentation in group-based samples is verified. And the pre-training in SGMC increases a little bit more time complexity, however this process contributes greatly to the improvement of model performance.

There are some limitations in this work. First, based on the group-based principle in SGMC, a kind of counterbalancing effect reduces the efficiency of sampling during the pre-training process. For each video clip, SGMC randomly selects only $2Q$ subjects to construct the group-based batch.

![Fig. 11](image)

**Fig. 11** The accuracies on DEAP (left) and SEED (right) with the number of epoch in the pre-training process of SGMC. $\text{acc}_\text{pre}$ represents percentage of correctly recognized positive pairs among all positive pairs, $\text{Epoch}_\text{pre}$ represents the number of epochs in the pre-training.

|                      | DEAP (s/epoch) | SEED (s/epoch) |
|----------------------|----------------|----------------|
| Proposed (Pre-training) | 11.762         | 30.363         |
| Proposed (Fine-tuning)   | 26.207         | 184.070        |
| Proposed (Fully-supervised) | 25.416        | 184.742        |

© Springer
in each epoch. This procedure may lead to the pre-training not being equally distributed in the samples. Therefore, the model has to increase the number of pre-training epochs to ensure that each sample is fully trained, and the calculation cost is greatly increased. It is necessary to design a new method with the group-based principle to improve the sampling efficiency in future work. For example, the first in first out (FIFO) method prioritizes sampling with fewer sampling times, so that all samples can be evenly and fully trained with fewer epochs. Second, it is worth further investigating the relationship between the learned representation in SGMC and the specific brain function for emotion recognition. Our visualization results explain that the learned representation is distinguishable not only in different emotions, but also in different video stimuli. However, SGMC mixes all channel information to learn the representation and cannot obtain brain topographic maps to understand the functional impact of specific brain areas. In future work, SGMC can be further improved by keeping the single channel information or introducing a new method to interpret the relationship between the features and brain function for emotion recognition.

Author Contributions Haoning Kan, Jiajing Huang and Haiyan Zhou designed and developed the model. Haoning Kan, Jiale Yu, Jiajing Huang and Zihe Liu performed the experiments and analyzed the data. Haoning Kan, Jiale Yu, Heqiang Wang and Haiyan Zhou wrote the article.

Funding The work is supported by Beijing Natural Science Foundation (No. 4222022).

Availability of data and materials All data generated or analyzed during this study are included in this article.

Code Availability The code is provided publicly online on https://github.com/kanhaoning/Self-supervised-group-meiosis-contrastive-learning-for-EEG-based-emotion-recognition.

Declarations

Conflicts of interest The authors declared that they have no conflicts of interest to this study.

References

1. Kalagginis FP, Adams D, Laskaris N (2018) Musical neuropicks: a consumer-grade bci for on-demand music streaming services. Neurocomputing 280:65–75
2. Pandey P, Swarnkar R, Kakaria S et al (2019) Understanding consumer preferences for movie trailers from EEG using machine learning. In: The 6th Annual Conference of Cognitive Science, pp 98–101
3. Subasi A, Tuncer T, Dogan S et al (2021) EEG-based emotion recognition using tunable q wavelet transform and rotation forest ensemble classifier. Biomed Signal Process Control 68(102):648
4. Tuncer T, Dogan S, Baygin M et al (2021) Tetromino pattern based accurate EEG emotion classification model. Artif Intell Med 123(102):210
5. Dogan A, Barua PD, Baygin M et al (2022) Automated accurate emotion classification using clefia pattern-based features with EEG signals. Int J Healthc Manag. https://doi.org/10.1080/20479700.2022.2141694
6. Zhang X, Yao L, Wang X et al (2019) A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. J Neural Eng. https://doi.org/10.1088/1741-2552/ABC902
7. Song T, Zheng W, Song P et al (2020) EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Trans Affect Comput 11:532–541
8. Du X, Ma C, Zhang G et al (2022) An efficient lstm network for emotion recognition from multichannel EEG signals. IEEE Trans Affect Comput 13:1528–1540
9. Tao W, Li C, Song R et al (2020) EEG-based emotion recognition via channel-wise attention and self attention. IEEE Trans Affect Comput 14(1):382–393
10. Becker H, Fleurteau J, Guillotet P et al (2020) Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources. IEEE Trans Affect Comput 11:244–257
11. Zhang G, Yu M, Liu Y et al (2021) SparseDGCNN: Recognizing emotion from multichannel EEG signals. IEEE Trans Affect Comput 14(1):537–548
12. Wang Z, Zhang JW, He Y et al (2022) EEG emotion recognition using multichannel weighted multiscale permutation entropy. Appl Intell 52:12.064–12.076
13. Ye Z, Xie X, Liu Y et al (2021) Towards a better understanding of human reading comprehension with brain signals. The ACM Web Conference 2022:380–391
14. Zhang X, Zhao Z, Tsiligkaridis T et al (2022) Self-supervised contrastive pre-training for time series via time-frequency consistency. In: 2022 Neural Information Processing Systems (NeurIPS)
15. Banville HJ, Chehab O, Hyvärinen A et al (2020) Uncovering the structure of clinical EEG signals with self-supervised learning. J Neural Eng 18(4):46,020
16. Jiang X, Zhao J, Du B et al (2021) Self-supervised contrastive learning for EEG-based sleep staging. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp 1–8
17. Dmochowski JP, Bezdek MA, Abelso B et al (2014) Audience preferences are predicted by temporal reliability of neural processing. Nat Commun 5(1):4567
18. Dmochowski JP, Sajda P, Dias J et al (2012) Correlated components of ongoing EEG point to emotionally laden attention - a possible marker of engagement? Front Hum Neurosci 6:112
19. Shen X, Liu X, Hu X et al (2022) Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2022.3164516
20. Ma Y, Zhao W, Meng M et al (2023) Cross-subject emotion recognition based on domain similarity of EEG signal transfer learning. IEEE Trans Neural Syst Rehabilitation Eng 31:936–943
21. Tian C, Ma Y, Cammon J et al (2023) Dual-encoder vae-gan with spatiotemporal features for emotional EEG data augmentation. IEEE Trans Neural Syst Rehabilitation Eng 31:2018–2027
22. Tuncer T, Dogan S, Subasi A (2021) Ledpatnet19: Automated emotion recognition model based on nonlinear led pattern feature extraction function using EEG signals. Cogn Neurodyn 16:779–790
23. Peng Y, Liu H, Li J et al (2022) Cross-session emotion recognition by joint label-common and label-specific EEG features exploration. IEEE Trans Neural Syst Rehabilitation Eng 31:759–768
24. Alhagry S, Fahmy AA, El-Khoribi RA (2017) Emotion recognition based on EEG using LSTM recurrent neural network. Int J Adv Comput Sci Appl 8(10):355–358
25. Li Y, Wang L, Zheng W et al (2021) A novel bi-hemispheric discrepancy model for EEG emotion recognition. IEEE Trans Cogn Dev Syst 13:354–367
26. Tripathi S, Acharya SG, Sharma RD et al (2017) Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. In: 2017 AAAI, pp 4746–4752
27. Salama ES, El-Khoribi RA, Shoman M et al (2018) EEG-based emotion recognition using 3D convolutional neural networks. Int J Adv Comput Sci Appl 9(8):329–337
28. Yang Y, Wu Q, Qu M et al (2018) Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp 1–7
29. Maria MA, Akhand MAH, Hossain ABMA et al (2023) A comparative study on prominent connectivity features for emotion recognition from EEG. IEEE Access 11:37,809–37,831
30. Cheah KH, Nisar H, Yap VV et al (2021) Optimizing residual networks and VGG for classification of EEG signals: identifying ideal channels for emotion recognition. J Healthc Eng. https://doi.org/10.1155/2021/5599615
31. Gu Y, Zhong X, Qu C et al (2023) A domain generative graph network for EEG-based emotion recognition. IEEE J Biomed Health Inform 27:2377–2386
32. Niu W, Ma C, Sun X et al (2023) A brain network analysis-based double way deep neural network for emotion recognition. IEEE Trans Neural Syst Rehabilitation Eng 31:917–925
33. Lu W, Tan TP, Ma H (2023) Bi-branch vision transformer network for EEG emotion recognition. IEEE Access 11:36,233–36,243
34. Song Y, Zheng Q, Liu B et al (2022) EEG conformer: Convolutional transformer for EEG decoding and visualization. IEEE Trans Neural Syst Rehabilitation Eng 31:710–719
35. Gidaris S, Singh P, Komodakis N (2018) Unsupervised representation learning by predicting image rotations. In: 2018 International Conference on Learning Representations (ICLR)
36. Sermanet P, Lynch C, Chebotar Y et al (2018) Time-contrastive networks: Self-supervised learning from video. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp 1134–1141
37. Mikolov T, Sutskever I, Chen K et al (2013) Distributed representations of words and phrases and their compositionality. In: 2013 Conference and Workshop on Neural Information Processing Systems (NIPS), pp 3113–3119
38. Devlin J, Chang MW, Lee K et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: 2019 North American Association for Computational Linguistics (NAACL), pp 4171–4186
39. Saeed A, Salim FD, Ozcelebi T et al (2021) Federated self-supervised learning of multisensor representations for embedded intelligence. IEEE Internet Things J 8:1030–1040
40. Sarkar P, Etemad A (2020) Self-supervised learning for ecg-based emotion recognition. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 3217–3221
41. Zheng Z, Feng X, Yu H et al (2022) Unsupervised few-shot image classification via one-vs-all contrastive learning. Appl Intell 53(7):7833–7847
42. Zhang Z, Zhong Sh, Liu Y (2022) GANSER: a self-supervised data augmentation framework for EEG-based emotion recognition. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2022.3170369
43. Qi C, Su H, Mo K et al (2017) PointNet: deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 77–85
44. Chen T, Kornbluh S, Norouzi M et al (2020) A simple framework for contrastive learning of visual representations. In: The 37th International Conference on Machine Learning (ICML), pp 1597–1607
45. Paszke A, Gross S, Massa F et al (2019) Pytorch: an imperative style, high-performance deep learning library. In: 2019 Conference and Workshop on Neural Information Processing Systems (NeurIPS), pp 8024–9035
46. Kingma D, Ba J (2015) Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR)
47. Ma JX, Tang H, Zheng WL et al (2019) Emotion recognition using multimodal residual LSTM network. In: The 27th ACM International Conference on Multimedia, pp 176–183
48. Gao Z, Wang X, Yang Y et al (2021) A channel-fused dense convolutional network for EEG-based emotion recognition. IEEE Trans Cogn Dev Syst 13:945–954
49. Tao W, Li C, Song R et al (2020) EEG-based emotion recognition via channel-wise attention and self attention. IEEE Trans Affect Comput 14(1):382–393
50. Dong Y, Ren F (2020) Multi-reservoirs EEG signal feature sensing and recognition method based on generative adversarial networks. Comput Commun 164:177–184
51. Li Y, Zheng W, Cui Z et al (2018) EEG emotion recognition based on graph regularized sparse linear regression. Neural Process Lett 49:555–571
52. Song T, Zheng W, Song P et al (2020) EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Trans Affect Comput 11:532–541
53. van der Maaten L, Hinton GE (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
54. Zhang H, Cisse M, Dauphin YN et al (2018) Mixup: beyond empirical risk minimization. In: 2018 International Conference on Learning Representations (ICLR)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.
Authors and Affiliations

Haoning Kan¹ · Jiale Yu² · Jiajin Huang³ · Zihe Liu¹ · Heqian Wang¹ · Haiyan Zhou³

Haoning Kan
KanHaoning@emails.bjut.edu.cn

Jiale Yu
jiale.yu@ucdconnect.ie

Jiajin Huang
jhuang@bjut.edu.cn

Zihe Liu
ZiheLiu@emails.bjut.edu.cn

Heqian Wang
vbfncac@emails.bjut.edu.cn

¹ Faculty of Science, Beijing University of Technology, 100 Pingleyuan, 100124 Beijing, China

² Beijing-Dublin International College, Beijing University of Technology, 100 Pingleyuan, 100124 Beijing, China

³ Faculty of Information Technology, Beijing University of Technology, 100 Pingleyuan, 100124 Beijing, China