Self-Supervised EEG Emotion Recognition Models Based on CNN

Xingyi Wang, Yuliang Ma, Jared Cammon, Feng Fang, Yunyuan Gao, and Yingchun Zhang

***Abstract*— Emotion plays crucial roles in human life. Recently, emotion classification from electroencephalogram (EEG) signal has attracted attention by researchers due to the rapid development of brain computer interface (BCI) techniques and machine learning algorithms. However, recent studies on emotion classification show resource utilization because they use the fully-supervised learning methods. Therefore, in this study, we applied the self-supervised learning methods to improve the efficiency of resources usage. We employed a self-supervised approach to train deep multi-task convolutional neural network (CNN) for EEG-based emotion classification. First, six signal transformations were performed on unlabeled EEG data to construct the pretext task. Second, a multi-task CNN was used to perform signal transformation recognition on the transformed signals together with the original signals. After the signal transformation recognition network was trained, the convolutional layer network was frozen and the fully connected layer was reconstructed as emotion recognition network. Finally, the EEG data with affective labels were used to train the emotion recognition network to clarify the emotion. In this paper, we conduct extensive experiments from the data scaling perspective using the SEED, DEAP affective dataset. Results showed that the self-supervised learning methods can learn the internal representation of data and save computation time compared to the fully-supervised learning methods. In conclusion, our study suggests that the self-supervised machine learning model can improve the performance of emotion classification compared to the conventional fully supervised model.**

***Index Terms*— EEG, self-supervised, emotion classification, multi-task learning.**

I. Introduction

EMOTIONS, as expressions of human physiology and psychology, is related to personality, preferences, and physical and mental states [1]. Emotion recognition has been widely studied in many areas. For example, previous study has reported that emotion recognition can be utilized to assess driver's physical and mental conditions and identify whether the driver is focused or distracted to avoid traffic accidents [2]. Besides, emotion recognition has been demonstrated to be able to identify people's negative emotions and decrease the chance to get diseases such as depression [3]. Emotion recognition has also been used in human-robot interaction to enable service-oriented robots to understand human emotions and communicate with people [4]. Recently, there are two main approaches to recognize different emotions. The first one is through interpreting behavioral signals, such as motor gait [5], facial expressions [6], speech [7], and text [8]. These signals are susceptible to external interference and easily induce noises during signal acquisition, resulting low recognition accuracy. The second type is through interpreting physiological signals from the brain, heart, and eyes using electroencephalogram (EEG) [9], electrocardiogram (ECG) [10], electrooculogram(EOG) [11], galvanic skin response (GSR) [12] and so on. These signals are spontaneously generated, not easily controlled, and therefore have been widely studied in emotion recognition. In particular, brain signals recorded from EEG have been widely studied as the focus of emotion recognition [13], [14]. For example, previous study employed EEG to identify different emotions and demonstrated that there is a correlation between emotional states, neural activity regions of the brain, and EEG signal frequency bands [15].

Recently, EEG-based machine learning algorithms have been developed to recognize various emotions. Current researches on EEG-based emotion recognition has generally used the fully-supervised machine learning model. For example, in [16], the authors applied improved HCNN (hierarchical convolutional neural networks) to EEG classification. However, supervised models have some limitations. First, supervised learning requires a large amount of labeled data to train the model. However, the sample size of EEG data is usually limited due to the difficulty in recruiting a large cohort of subjects and the time consuming in collecting and labelling the EEG data. The relatively small amount of labeled data results in an undertrained model. In addition, the relatively large amount of unlabeled data is wasted since it cannot be used in supervised learning. Moreover, the supervised model is limited to a certain class of task and shows poor performance when applied to other tasks. Differently, self-supervised model overcomes these limitations by using a pre-training step on unlabeled data to initialize a deep learning model

Manuscript received 10 January 2023; revised 14 February 2023; accepted 13 March 2023. Date of publication 31 March 2023; date of current version 11 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62071161, Grant 61971168, and Grant 61372023; and in part by the Graduate Research Innovation Fund of Hangzhou Dianzi University under Grant CXJJ2022146. *(Corresponding authors: Yuliang Ma; Yingchun Zhang.)*

Xingyi Wang, Yuliang Ma, and Yunyuan Gao are with the School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: 212060347@hdu.edu.cn; mayuliang@hdu.edu.cn; gyy@hdu.edu.cn).

Jared Cammon, Feng Fang, and Yingchun Zhang are with the Department of Biomedical Engineering, University of Houston, Houston, TX 77204 USA (e-mail: jacammon@cougarnet.uh.edu; ranjitfeng@gmail.com; yingchun.umn@gmail.com).

Digital Object Identifier 10.1109/TNSRE.2023.3263570

(Self-supervised automatically generates labels by constructing pretext tasks), followed by fine-tuning on a labeled training set and evaluating on the corresponding test set [17]. Thus, self-supervised learning can use unlabeled data to learn a more comprehensive representation of the underlying structure of the data [18]. Since self-supervision learns a representation inherent to the data, it can be fine-tuned to different tasks. There is no need to train the network from scratch for different classification tasks, which is a feature that greatly improves performance of the network. Previous study has shown that self-supervised methods can improve the robustness of models and reduce their uncertainty [19].

Yet, there is a shortage of self-supervised models based on convolutional neural network (CNN) in EEG emotion recognition [17]. Inspired by other papers, this paper introduces self-supervised method in EEG emotion recognition and investigates EEG emotion recognition based on self-supervised CNN from the perspective of data scalability. There are mainly three research goals in this paper:

(1) Introducing self-supervised method in emotion recognition. The self-supervised training has two stages: At the first stage, six signal transformations are constructed for the unlabeled EEG signals and the corresponding labels are generated. The transformed signals with the original signals and the corresponding labels are trained by a self-supervised multi-task CNN. Next, the convolutional layers of the trained network are frozen and the corresponding dense layers are added as the emotion recognition network. The network is then retrained using the EEG signals labeled with affective labels.

(2) Exploring the effect of data feature on the self-supervised model. In the SEED dataset, the original data and the data after extracting DE(Differential Entropy) features are separately used for self-supervised training to observe the effect of data feature on the pretext tasks as well as the downstream tasks.

(3) Investigating the effect of the volume of the dataset on the self-supervised model. In the DEAP dataset, on the basis of the extracting DE feature data, the self-supervised pre-training uses 20%, 30%, 40%, 50%, 60%, 70%, and 80% of the training sets, respectively. The Russell model is used to transform the valence and arousal labels of the dataset into 8 emotion labels, the trained signal transformation recognition network is fine-tuned to the valence dichotomous classification task, arousal dichotomous classification task, and 8-emotion classification task simultaneously.

The experimental results show that the self-supervised learning is able to identify the effective labels well on both SEED and DEAP datasets. On the SEED dataset, the average accuracy of emotion classification is 84.54% for SEED preprocessed data and 98.65% for SEED DE feature data. On the DEAP dataset, the signal transformation recognition network is migrated to three different classification tasks and trains with different amounts of data, and the three different tasks achieve good results on the affective classification. From the perspective of data scaling, it is demonstrated that the extracted feature is better for self-supervised multi-task learning than the original data; it is confirmed that the signal transformation recognition can learn the intrinsic features of the data. Meanwhile, the self-supervised model is compared with other supervised methods, and the results show that the self-supervised model outperforms the fully-supervised model, and can be extended to multiple tasks in the same domain(in this paper, same domain refers to the same dataset), improving the performance of the network and saves computation time.

II. Related Work

A. EEG-Based Affective Computing

Many studies have demonstrated that using EEG signals and machine learning algorithms for emotion recognition is reliable. In prior study [20], the authors explored the effects of EEG signal frequency bands in emotion classification using the discrete wavelet transform to decompose the EEG signal into $\gamma, \beta, \alpha, \theta$ frequency bands, and extracting spectral features from each frequency band. SVM(Support Vector Machines) was used to recognize emotion on the DEAP dataset and obtain the highest classification accuracy of spectral features in the β band. The arousal classification accuracy was 91.3%, and the valence classification accuracy was 91.1%. Previous study also explored the relationship between emotion classification and different features by extracting power, energy, differential entropy, and time-frequency features from five frequency bands of EEG signals, and used SVM to classify emotions on the SEED-IV dataset [21]. The authors concluded that the four emotion classification accuracies were 79%, 76%, 77% and 74%, respectively. Although machine learning can complete the classification task, it requires manual extraction of features during the classification process. Deep learning then solves this problem. In the literature [22], the authors used CNN to classify emotions from EEG signal with extracted DE features, using the SEED dataset for validation, achieving an average accuracy of 90.41%. In the literature [23], the authors explored the relationship between electrode positions and the performance of emotion recognition by treating all channel features as a three dimensional feature matrix and processing the three dimensional feature matrix using CNN, achieving 85.88% accuracy in valence classification and 85.53% accuracy in arousal classification on the DEAP dataset.

B. Self-Supervised Learning

Self-supervised learning has been widely used in many fields. In natural language processing, self-supervised models like GPT(General Pre-Training), BERT(Bidirectional Encoder Representation from Transformers), can be well applied to tasks such as machine translation and language modeling [24]. In computer vision, self-supervised models like SimCLR can be applied to image classification tasks [25]. In the literature [28], the author used multi-task convolutional neural network for self-supervised learning, learned the features of pre-processed data through multi-task learning, and studied human activity detection. Acknowledging the validity of self-supervision, we apply it to emotion recognition of physiological signals. In the literature [17], the authors constructed seven pretext tasks for self-supervised training based on ECG signals, after which the models were received to downstream

tasks for emotion recognition studies. In the literature [26], the authors used contrast learning to study sleep EEG features, constructed positive and negative samples for self-supervised learning. In the literature [27], the authors also studied self-supervision on Transformer architecture for EEG emotion recognition, The authors proposed graph-based multi-task self-supervision (GMSS) to learn EEG emotion representation, as well as evaluated GMSS on SEED, SEED-IV.

III. Method

Self-supervised learning is trained on unlabeled data so that the model learns the intrinsic representations of the data and then fine-tunes to downstream tasks. Self-supervised learning generally has two steps: first, the corresponding labels are generated on the unlabeled data by constructing pretext tasks. The network is then trained using the generated label information. This part is to perform the signal transformation recognition task, which is defined as T_p After that, the learned model's convolutional layer is frozen, the dense layer is reconstructed and trained using the label information from the downstream task. This part is performed by the emotion recognition task, which is defined as T_d. In this paper, we study the self-supervised CNN-based emotion recognition. The self-supervised method has two networks, including signal transformation recognition network and emotion recognition network. These two networks are described below.

A. Signal Transformation Recognition Network

The features of a single task cannot achieve good results for classification of multiple tasks. In order to learn the universal features among multiple tasks and increase the model performance, a multi-task CNN is used for the signal transformation recognition network. The multi-task CNN contains two parts, the shared layer and the task-specific layer. The shared layer, as the low level of the neural network, is able to learn the universal features of the data. This makes the model have better generalization performance by learning the universal features of the data among multiple tasks, so only the shared layer is migrated for the emotion recognition network later. The task-specific layer, as the high level of the neural network, is able to learn specific feature information to classify pretext tasks.

The network model in this paper has three convolutional blocks as shared layers, and each shared layer consists of two one-dimensional convolutional layers, which are ReLu activation function, and the BN layer. The size of the convolutional kernels is reduced from 32 to 16 and 8, the number of convolutional kernels is increased from 32 to 64 and 128 [17], [28], and the BN layer can be selected according to specific data. A pooling operation is performed after each convolutional block, using a pooling layer of size 8. A global maximum pooling operation is performed after the last convolutional block, and the output is sent to a task-specific layer through a fully connected layer. The task-specific layer generates a branch for each signal transformation, each branch consisting of two dense layers and the ReLu activation function, while a 60% droupout is introduced to prevent overfitting (droupout is not used when using the BN layer). The signal transformation recognition model is shown in Fig 1.

1) Pretext Tasks: In order to make self-supervised learning fully learn the characteristics of the data such as time, noise and other dimensions, this paper performs six signal transformations on the original EEG signal as pretext tasks, and the six transformed signals are labeled with corresponding labels together with the original signal and sent to the signal transformation recognition network as model input. The six signal transformations are described below.

(1) Adding noise: random noise with Gaussian distribution, N(t), added to the original EEG signal, S(t).

(2) Scale transformation: the amplitude of the original EEG signal is stretched or telescoped. The amplitude of original EEG signal, S(t), is transformed into α^*S(t), and α is the scale factor.

(3) Signal horizontal flipping: the original EEG signal is flipped according to the horizontal line. The amplitude of original EEG signal, S(t), is transformed into -1^*S(t).

(4) Signal vertical flipping: the original EEG signal is flipped according to the vertical line (actually, a time flip is performed).

(5) Temporal dislocation: the original EEG signal is evenly divided into n segments, randomly select the seed numbers and recombined in a random order to achieve perturbation of the temporal position of each segment.

(6) Time warping: the original signal is evenly divided into m segments, randomly select the seed numbers to stretch or compress each segment in time, recombine each segment of the signal after time warping, and later construct the data dimension to the same dimension as the original data.

After performing the 6 signal transformations, the original signal is stacked with the 6 signal transformations to generate corresponding labels [0,1,2,3,4,5,6]. The 7 signals and corresponding labels are the inputs of the signal transformation recognition network. For different datasets and different input signals, the parameters of the 6 transformation tasks should be fine-tuned accordingly, and the parameter adjustment is mainly based on experience. The detailed parameter settings of the pretext tasks are shown in Table I.

2) Loss Function: Since the transformed signal labels are independent of each other and the network classifies multiple labels, the network uses sigmoid-activated cross-entropy loss function for the classification loss function.

The input of transformation recognition task T_p is defined as the tuple table (X_j, y_j) where X_j is the j transformed signal, the y_j is the corresponding label generated by the j_{th} transformation, and $j \in [0, N]$, where N is the total number of transformations of the signal, which is equal to $L*7$, where L is the total length of the data. Accurate classification of each task is achieved by minimizing the cross-entropy loss function. The loss function of each label is defined by the following equation, where the prediction probability of the j_{th} task is defined as P_j:

$$L_j = -[y_j ln P_j + (1 - y_j) ln(1 - P_j)] \tag{1}$$

In order to learn the signal transformation recognition task T_p, the final total loss consists of the weighted value of the loss

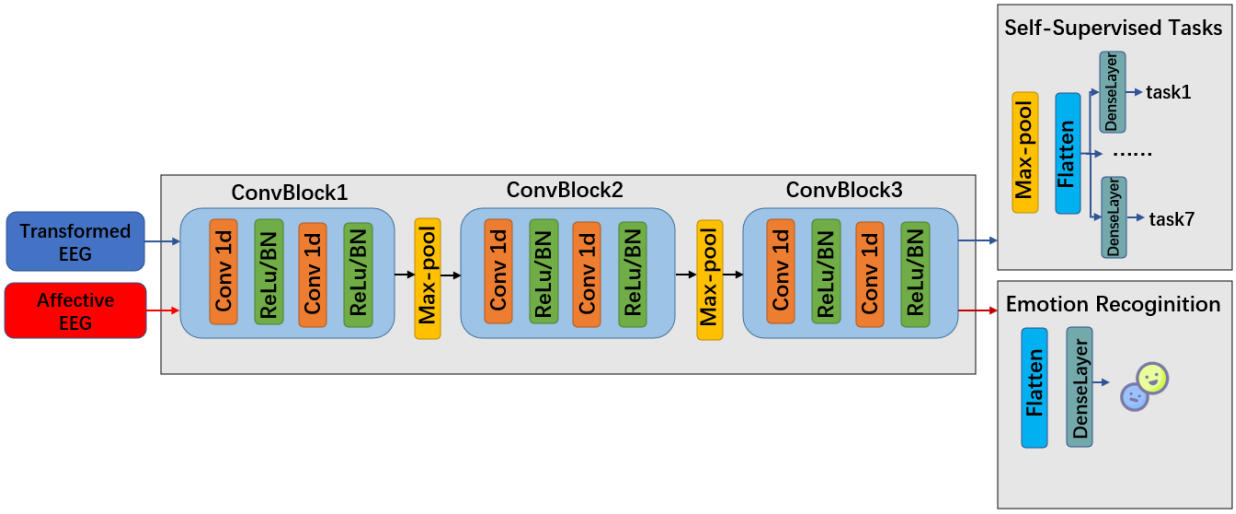

Fig. 1. The figure shows the self-supervised architecture. First, the transformed EEG data is used for learning EEG representations by self-supervised learning. Then, the convolution layers are transferred to the emotion recognition network, where the fully connected layers are trained to classify emotions.

TABLE I
PRETEXT TASK PARAMETER SETTINGS

Task Transformation	SEED Original	SEED Features	DEAP Features
Amount of noise	15	15	15
Scale factor	1.2	1.2	1.2
Time dislocation pieces	10	5	4
Time distortion pieces	10	5	8
Time distortion stretch factor	1.05	1.05	1.05
Time distortion compression factor	0.95	0.95	0.95

of each pretext task. The total loss of the network is defined as L_{total}, and the loss factor is defined as β_j, and the total loss is defined by the equation:

$$L_{total} = \sum_{j=0}^{N} \beta_j L_j \tag{2}$$

B. Emotion Recognition Network

The emotion recognition network includes the convolutional layers for signal transformation recognition network, and the fully connected layers with hidden nodes.

The trained signal transformation recognition network has learned the universal features of EEG data. The convolutional layers of the trained signal transformation recognition network is frozen and migrated, after which different fully connected layers are constructed according to different datasets, and the fully connected layers are retrained using EEG data with affective labels, so that the high level of the network learns the features of the emotion classification task.

1) Loss Function: The input of emotion recognition task T_d is defined as the tuple table (X_i, y_i), where X_i is the original EEG signal, and y_i is the corresponding emotion label, the $i \in [0, M]$, where M is the number of EEG input signals. The accurate classification of each emotion is achieved by minimizing the cross-entropy loss function, which is defined by the following equation:

$$L_i = \sum_{i=1}^{M} y_i \ln q_i \tag{3}$$

where the prediction probability of the i_{th} task is defined as q_i.

2) Fully-Connected Layer Parameters: The fully connected layers of the emotion recognition network is not migrated from the signal transformation recognition network, but is reconstructed and retrained using EEG data with emotion labels. To evaluate the ability of the self-supervised method to learn robust generalized EEG, a simple single-task fully connected layer with a shallow network layer is constructed. Specifically, four dense layers with L2 regularization and 40% droupout are constructed for the SEED preprocessed data; two dense layers with L2 regularization and 40% droupout and 20% droupout are constructed for the SEED and DEAP DE features. The final output comes from sigmoid (binary classification) or softmax(multi classification) layers. In this paper, the parameters of the dense layers are fine-tuned for different classification tasks for each dataset. The number of hidden nodes was set according to the input of the model. Because there is a large amount of redundant information in the pre-processed data, the number of hidden nodes and the number of fully-connected layer were set relatively large, because the feature data has been filtered out a lot of redundant information, the number of fully-connected layer and the

TABLE II
FULL CONNECTED LAYER PARAMETER SETTINGS

Dataset	Dense layer parameters
SEED pre-processed	epoch: 300; batch size: 256; learning rate: 0.0001; hidden nodes: 640
SEED feature data	epoch: 200; batch size: 256; learning rate: 0.0001; hidden nodes: 192
DEAP feature data	epoch: 200; batch size: 256; learning rate: 0.0001; hidden nodes: 192

number of hidden nodes were set relatively small in order to prevent over-fitting. And in the DEAP dataset, in order to see how self -supervision affects the performance of different downstream tasks, the dense layer parameters of the three classification tasks are intentionally made the same. The parameters such as epoch, batch size, learning rate were set by referring to literatures [17] and in order to make the model converge, the parameters were adjusted accordingly in the process of training the model. The detailed parameter settings are shown in Table II.

IV. DATASET AND DATA PROCESSING

A. Dataset

In this paper, experiments are conducted using two international public datasets, SEED [29], [30] and DEAP [31], to investigate the effect of data features on the self-supervision effect on the SEED dataset and explore the effect of data volume on the self-supervision effect on the DEAP dataset. The two datasets and the data processing are described below.

1) SEED Dataset: The SEED dataset collected experimental data from 15 subjects (7 males and 8 females). During the experiment, each subject watched 15 movie clips (5 positive clips, 5 negative clips, and 5 neutral clips), for a total of 15 trails. During a trail, the movie had a 5s cue, the movie ran for 4 minutes, self-assessment was 45s, and the break was 15s. Each volunteer performed 3 experiments, each experiment was separated by 1 week, for a total of 45 experimental data. The SEED dataset had 3 categories of emotions: positive, negative, and neutral. In this paper, the goal of the SEED dataset is 1 triple classification problem.

2) DEAP Dataset: The DEAP dataset collected experimental data from 32 subjects (16 males and 16 females). During the experiment, each subject watched 40 music videos. During a music video, each clip was preceded by a 2s cue, a 5s baseline recording, 63s music video playback time, and 15s self-assessment.

The DEAP dataset has three sentiment evaluations, namely: valence, arousal, and dominance, which are represented by integers from 1-9 as state levels. In this paper, the sentiment evaluation of the DEAP dataset is processed in three ways. First, the emotion expressions in two dimensions of valence and arousal into eight emotion categories according to Russell's model. Second, the state levels of valence into binary, with 5 as the threshold, and those above 5 as high valence and those below 5 as low valence. Third, the state levels of arousal into binary, with the same principle as valence. In this paper, the experimental objectives of the DEAP dataset are two binary classification problems and one 8 classification problem.

TABLE III
DATA FORMAT DESCRIPTION TABLE

Dataset	Data Format
SEED preprocessed data	(62 channels*50 sampling points) *1
SEED DE feature data	(62 channels*5 bands) *1
DEAP DE feature data	(32 channels*4 bands) *1

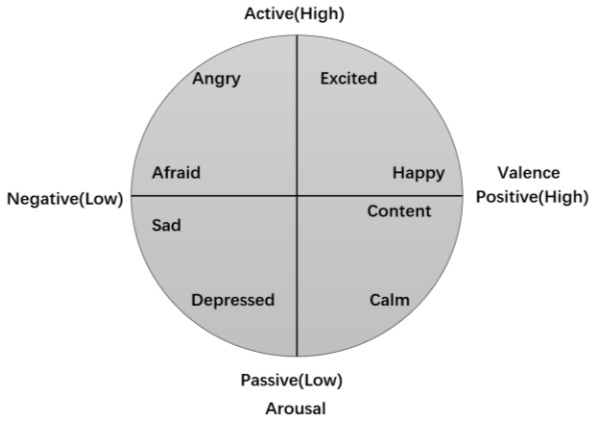

Fig. 2. Russell's emotion model.

B. Data Processing

The model in this paper uses one-dimensional convolution, so eventually all the data have to be processed into M*1 format. For the SEED raw data, the normalization process is done in the time dimension, after the raw data of each subject is cut into 3100*1 signal. Since the feature data is provided in the SEED dataset, there is no need to perform feature extraction separately. On the feature data provided in the dataset, the normalization process is done in the feature dimension, and then the feature data of each subject is cut into 310*1 signal. For the DEAP dataset, first the DE features are extracted, after the normalization process is done in the feature dimension, and finally the feature data of each subject is cut into 128*1 signal. The data formats of the three kinds of data are shown in Table III.

1) Emotion Model: The psychologist Russell categorized emotions by introducing "dimensions". He proposed a two-dimensional circular model, in which emotions can be classified into two latitudes: pleasantness and intensity. In this paper, the valence and arousal of the DEAP are transformed into eight emotions for classification according to Russell's model. The emotion model is shown in Fig 2.

2) DE Characteristic: DE features are more suitable for emotion recognition than other features, and the more

TABLE IV
F1 SCORES FOR SIGNAL TRANSFORMATION OF SEED PREPROCESSED DATA

f1_score							
Subjects	Signal Transformation						
	Original signal	Adding Noise	Scale change	Horizontal Flip	Vertical Flip	Time Dis-location	Time Warp
1	0.95	0.99	0.95	0.99	1.00	0.97	1.00
2	0.48	0.95	0.49	0.80	0.99	0.92	1.00
3	0.65	0.97	0.55	0.92	1.00	0.90	1.00
4	0.62	0.78	0.50	0.94	0.98	0.92	1.00
5	0.61	0.85	0.70	0.87	0.92	0.88	0.99
6	0.21	0.93	0.60	0.72	1.00	0.98	1.00
7	0.46	0.98	0.69	0.94	1.00	0.90	1.00
8	0.50	0.98	0.74	0.88	1.00	0.85	0.96
9	0.50	0.99	0.74	0.96	0.99	0.78	1.00
10	0.65	0.72	0.76	0.91	0.99	0.88	1.00
11	0.60	0.93	0.56	0.94	0.99	0.84	1.00
12	0.42	0.98	0.73	0.97	1.00	0.91	1.00
13	0.59	0.95	0.70	0.79	0.97	0.92	0.84
14	0.65	0.99	0.66	0.93	1.00	0.91	1.00
15	0.75	0.72	0.71	0.92	0.62	0.90	0.67
Average	0.5760	0.9140	0.6720	0.8987	0.9633	0.8973	0.9640
Std	0.1597	0.0945	0.1144	0.0725	0.0939	0.0475	0.0882

information based on context (the longer the duration), the more obvious the extracted DE features are. In this paper, when extracting DE features for the DEAP dataset, the signal is divided into multiple 3s segments according to the duration of the signal. In each segmentation, four frequency bands of the data are extracted (θ :4-8Hz, α :8-14Hz, β :14-31Hz, γ :31-45Hz) to calculate the DE features corresponding to each band.

V. EXPERIMENTAL RESULTS

A. Self-Supervised Comparison Experiments of SEED Data Features

In order to visualize the effect of data features on the self-supervised, both SEED preprocessed data and DE feature data are selected from 15 subjects for one experiment to train the signal transformation recognition network. In the process of training the signal recognition network, both 80% of the data are selected for the training set and 20% of the data are used for the test set. In this paper, F1 score is used as the evaluation index of the signal transformation recognition network, and the final classification results are shown in Table IV and Table V.

Table IV and Table V show the F1 scores of 15 subjects on the signal transformation recognition network. The results show that when detecting the original and scale-transformed signals on the preprocessed data, the average F1 scores are 57.60% and 67.20%, which are relatively low Vertical flip and time warping achieve very high scores on the F1 score, with multiple subjects reaching 1. On the DE feature data, all signal transformations achieve very high scores. In addition, the classification time used to train the signal transformation recognition network using the SEED feature data was reduced by 75% compared to the original SEED data, and the feature extraction effectively remove the redundant information from the data.

The average F1_score of the preprocessed data on the seven pretext tasks is 0.5760, 0.9140, 0.6720, 0.8987, 0.9633, 0.8973, 0.9640, respectively, and the average F1_score of the feature data on the seven pretext tasks all reach about 0.99. Due to the redundant information of the preprocessed data, it is obvious that the extracted features are better for signal transformation recognition than the preprocessed data, and the indicators on the signal transformation recognition network not only achieve good results but also achieve very low standard deviations.

B. SEED Emotion Classification Experiments

The self-supervised learning effect is reflected by the performance of the downstream tasks. During the training process of the emotion recognition network, the same training set and test set as the signal recognition network are selected for training and testing. In this paper, the accuracy, precision, recall, and F1_score are used as the evaluation index of the emotion recognition network, and the final classification results are shown in Figure 3.

Fig 3 respectively shows the contrast accuracy, precision, recall, and F1_score of 15 subjects on the emotion recognition

TABLE V
F1 SCORES FOR SIGNAL TRANSFORMATION OF SEED DE FEATURE DATA

Subjects	f1_score						
	Signal Transformation						
	Original signal	Adding Noise	Scale change	Horizontal Flip	Vertical Flip	Time Dis-location	Time Warp
1	0.99	1.00	1.00	1.00	1.00	0.99	1.00
2	0.99	1.00	1.00	1.00	1.00	0.99	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	0.99	1.00	0.99	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	0.99	1.00
7	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	0.99	1.00	1.00	1.00	1.00	0.99	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00	1.00	1.00	1.00
13	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	0.9973	1.00	0.9993	1.00	1.00	0.9973	1.00
Std	0.0044	0	0.0024	0	0	0.0044	0

network. The average accuracy, precision, recall, and F1_score of the preprocessed data are 0.8454, 0.8451, 0.8444, and 0.8447, respectively. The average accuracy, precision, recall, and F1_score of the feature data are 0.9865, 0.9867, 0.9865, and 0.9866, respectively. Table VI shows the standard deviation of each indicator for the 15 subjects. Due to the individual variability among different subjects, the emotion classification effect of the preprocessed data is not only lower in each index, but also the standard deviation among each subject is relatively large, and the data after secondary feature extraction improves the generalization ability of the model.

TABLE VI
THE STANDARD DEVIATION OF EACH INDICATOR

	accuracy	precision	recall	f1_score
Preprocessed Data	4.5932	4.5899	4.5898	4.5889
Feature data	0.8239	0.8185	0.7993	0.8092

C. DEAP Multi-Task Migration Experiment

To further verify the superiority of the self-supervised method, the DEAP dataset is also experimented with the emotion classification task. For emotion classification on the DEAP dataset, the trained signal transformation recognition network is migrated to three classification tasks simultaneously in this paper, namely, valence binary classification, arousal binary classification, and 8-emotion classification.

An experiment of 32 subjects from the DEAP dataset is selected and the DE feature is extracted to train the signal transform recognition network. In the eight-emotion classification experiments, because of the uneven distribution of the samples of emotion labels transformed using the Russell model, the training set and test set were divided according to the proportion of labels in each category. After data processing, there are 25,600 samples in the DEAP dataset. In this paper, 10%, 20%, 30%, 40%, 50%, 60%, 70%, and 80% of DEAP data are respectively selected as the training set, and all the remaining data are used as the test set to investigate the effect of data volume dimension on the emotion classification results and the effect of self-supervised learning on the migration of different classification tasks.

Table VII shows the F1 scores achieved by the signal transformation recognition network when different amounts of data are used as network inputs. The results show that when the data volume is 20% (5120 samples), all tasks of signal transformation recognition network can get high F1 score, and the larger the data volume, the higher the F1 scores obtained.

Fig 4 respectively shows the contrast accuracy, precision, recall and F1_score on the emotion recognition network with different data volumes. The results show that when the data volume reaches 20%, the valence and arousal classification metrics reach about 96%, and the accuracy increases gradually with the increase of data volume; when the data volume reaches 40%, the valence and arousal classification metrics

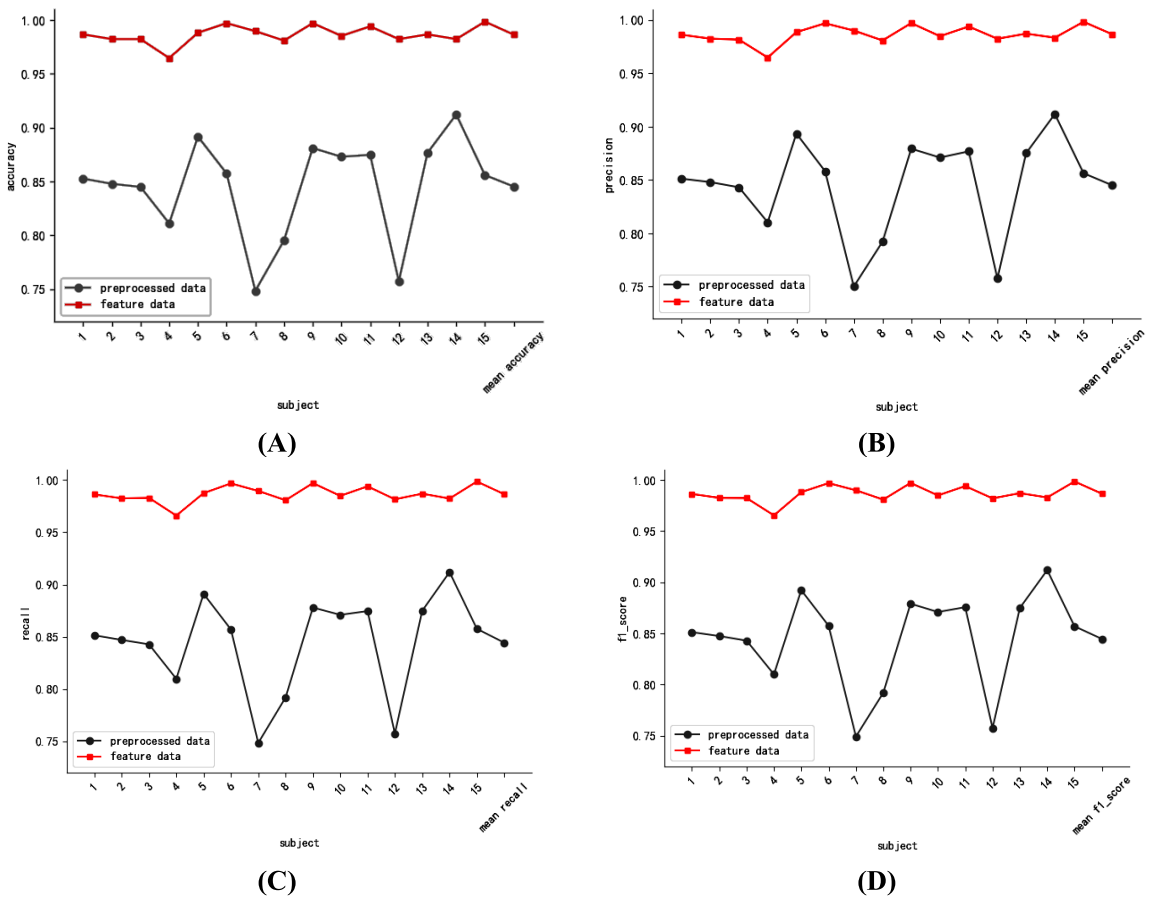

Fig. 3. The figure shows the contrast accuracy, precision, recall, and F1_score of 15 subjects between the processed data and the extracting DE feature data on the emotion recognition network.

TABLE VII
F1 SCORES OF DEAP FEATURE DATA SIGNAL TRANSFORMATION

Data volume	f1_score						
	Signal Transformation						
	Original signal	Adding Noise	Scale change	Horizontal Flip	Vertical Flip	Time Dis-location	Time Warp
10%	0.85	0.86	0.97	0.98	0.98	0.97	1.00
20%	0.93	0.93	0.99	0.99	1.00	0.98	1.00
30%	0.96	0.97	1.00	1.00	1.00	0.98	1.00
40%	0.94	0.95	0.98	0.99	1.00	0.98	0.99
50%	0.95	0.96	0.99	1.00	1.00	0.98	1.00
60%	0.96	0.98	1.00	1.00	1.00	0.98	1.00
70%	0.96	0.98	1.00	1.00	1.00	0.98	1.00
80%	0.96	0.98	1.00	1.00	1.00	0.98	1.00
Average	0.9388	0.9513	0.9913	0.9950	0.9975	0.9788	0.9988

reach about 99%. For the 8 emotion recognition, when the data volume reaches 60%, the indicators of the categories reach about 96%, and later with the increase of the data volume, the indicators basically stabilize at about 98%.

D. Baseline Experiment

Due to the lack of research on self- supervised in EEG emotion recognition. Therefore, I chose the emotion classification accuracy of some classical supervised algorithms to compare with the self-supervised algorithms. In the literature [32], the author extracted features from original EEG data and used a linear dynamic system approach to smooth these features. An average test accuracy of 87.53% was obtained by using all of the features together with a support vector machine in the SEED dataset; in the literature [15], the authors use a Discriminative Graph regularized Extreme learning machine

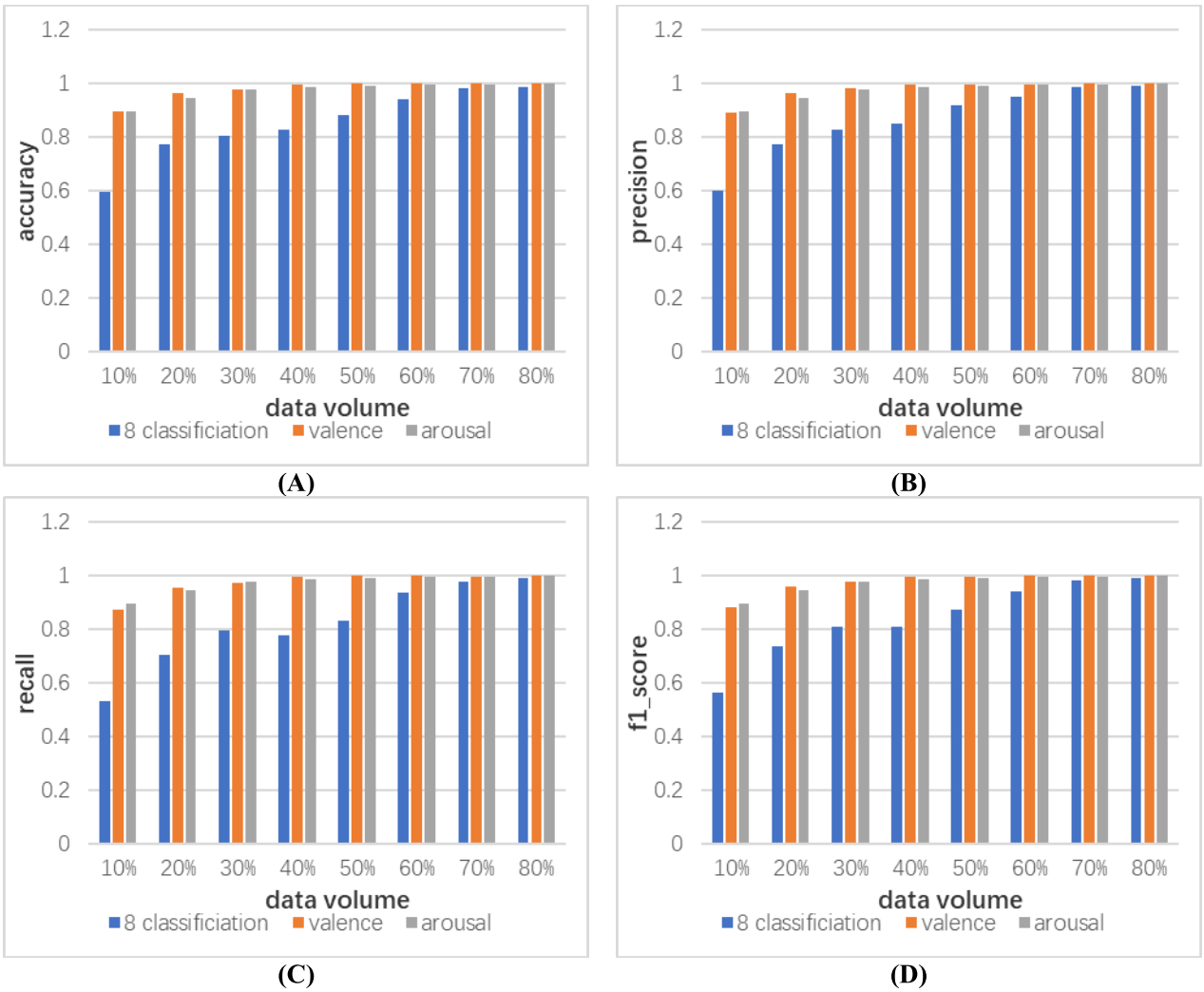

Fig. 4. The figure shows the contrast accuracy, precision, recall, and F1_score on the emotion recognition network with different data volumes.

with DE features achieves the accuracy of 91.07% for emotion classification on the SEED dataset; in the literature [33], the authors used a key channel and band detection method based on Deep Belief Network using differential entropy signals, and the final model had 86.08% classification accuracy in the SEED dataset; in the literature [34], the authors use support vector machines to classify the DEAP dataset with a final arousal classification accuracy of 64.90% and valence classification accuracy of 65.00%; in the literature [35], the authors propose a Temporal Convolutional Network and Broad Learning System, the DEAP dataset was used for the experiments, and the model achieved an average classification accuracy of 99.5755% and 99.5785% for valence and arousal, respectively. The results show that the effect of self-supervised algorithm is due to the effect of classical supervised algorithm. The contrast results are shown in TableVIII.

TABLE VIII
PERFORMANCE COMPARISON BETWEEN SELF-SUPERVISED ANDSUPERVISED ALGORITHM

Model	SEED	DEAP	
		Arousal	Valence
SVM	87.53%[32]	64.90%[34]	65.00%[34]
CNN	90.41%[22]	85.53%[23]	85.88%[23]
ELM	91.07%[15]	*	*
TCNBLS	*	99.57% [35]	99.57% [35]
DBN	86.08%[33]	*	*
Ours	98.65%	99.89%	99.73%

VI. DISCUSSION

In this study, we employed a self-supervised learning approach to train the classification model for emotion recognition based on EEG signals. The effects of data volume and data features on the self-supervised performance was investigated in terms of data scaling. We tested the efficacy of our model based on two publicly available dataset, which were DEEP and SEED datasets. Our results demonstrated that the self-supervised approach can significantly improve the classification performance compared to the fully supervised scheme. To our knowledge, this is the first time. application of multi-task convolutional neural networks trained using a self-supervised approach to EEG emotion recognition.

A. Data and Multitask Learning Relationships

We first investigated the relationship between data features and multi-task learning. Table IV shows that the preprocessed data achieved F1 scores of 57.6% and 67.2% on the two tasks

of original signal and scale transformation, respectively, and scores of the remaining tasks are around 90%, indicating that the task complexity for the identification of these two signal transformations for the preprocessed data is high. Table V shows that the DE feature data achieves high scores on all tasks, and the feature data achieves relatively stable scores on each subject with low standard deviations. Table VII also verifies that the feature data are better for multi-task learning. In Table VII, it can be seen that the data after extracting DE features, regardless of the amount of data, can achieve good scores on the signal transformation recognition network. It is confirmed that the data with extracted features are better for multi-task learning and the signal transformation recognition network is able to learn a wider range of features.

In the experiments with SEED preprocessed data, the downstream task and the signal transformation recognition network does not yield good results. In future studies, we may try to reduce the complexity of the pretext task or increase the length of the model to investigate whether it will improve the performance of the downstream task.

B. Multitask Learning With Downstream Tasks

The effect of multi-task learning is strongly related to the effect of downstream task learning. In the SEED preprocessing data, the F1 scores of individual pretext tasks are very low, which play an influence on the downstream tasks when migrating later, and the metrics of emotion recognition are not very high. In the SEED DE feature data, F1 scores of all pretext tasks are high and the metrics of emotion recognition are also high, indicating that the learning of the pretext tasks influences the effect of the downstream tasks.

In the DEAP dataset, Table VII shows that the signal transformation recognition network F1 scores are high when the amount of data is 20%, all of which reaches around 90%, and all of the metrics are around 77% when migrating the 8 classification task and 89% when migrating the 2 classification task. This indicates that the learning effect of the downstream task is not only related to the learning of the pretext task, but also to the learning difficulty of the downstream task itself.

C. Self-Supervised Learning With Downstream Tasks

In this paper, a multi-task migration study is conducted on the DEAP dataset, and the signal transformation recognition network is trained based on the DEAP dataset and is migrated to three different classification tasks at the same time. Figure 4 shows that for the binary classification experiments, when the training set sample is 20% of the dataset, all the metrics of the binary classification experiments reach about 98%, and the improvement of the metrics is smaller with the increase of the sample size. For the 8 classification experiments, when the training set sample is 70% of the dataset, all the metrics achieve 98%. This suggests that the 8-classification experiment is more difficult than the binary classification experiment, so more data volume is needed to learn a wider range of data features. When the data samples were sufficient, the same signal recognition network is migrated to three different classification tasks, and good results are achieved for all three tasks. It is proved that the self-supervised learning method is able to migrate the model to multiple tasks with good results, saving a lot of time and resources.

VII. Conclusion

In this paper, a self-supervised learning approach is introduced into an emotion recognition scheme for EEG signals. Two international public datasets, DEEP and SEED datasets, are utilized for emotion recognition. The impact of data volume as well as data features on the self-supervised performance is explored in terms of data scaling. The final results demonstrate that the self-supervised approach can significantly improve the classification performance compared to the fully supervised scheme. The network for the learning assistance task is able to learn the intrinsic features of the affective EEG data well. In practice, the self-supervised model can be migrated to multiple classified tasks within the same domain, which saves computing resources and time.

References

[1] T. Van Huynh, H.-J. Yang, G.-S. Lee, S.-H. Kim, and I.-S. Na, "Emotion recognition by integrating eye movement analysis and facial expression model," in *Proc. 3rd Int. Conf. Mach. Learn. Soft Comput.*, Jan. 2019, pp. 173–176.

[2] D. S. Moschona, "An affective service based on multi-modal emotion recognition, using EEG enabled emotion tracking and speech emotion recognition," in *Proc. IEEE Int. Conf. Consum. Electron. Asia (ICCE-Asia)*, Nov. 2020, pp. 1–3.

[3] W. Wu, M. Wu, and K. Yu, "Climate and weather: Inspecting depression detection via emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6262–6266.

[4] S. Alam, B. Johnston, J. Vitale, and M.-A. Williams, "Would you trust a robot with your mental health? The interaction of emotion and logic in persuasive backfiring," in *Proc. 30th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2021, pp. 384–391.

[5] S. A. Etemad and A. Arya, "Classification and translation of style and affect in human motion using RBF neural networks," *Neurocomputing*, vol. 129, pp. 585–595, Apr. 2014.

[6] M. H. Black et al., "Mechanisms of facial emotion recognition in autism spectrum disorders: Insights from eye tracking and electroencephalography," *Neurosci. Biobehav. Rev.*, vol. 80, pp. 488–515, Sep. 2017.

[7] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, Feb. 2021.

[8] M. Sridevi, "Text emotion classification using stacked CNNs and LSTMs," in *Proc. Int. Conf. Automat., Signal Process., Instrum. Control*, 2021, pp. 717–724.

[9] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul. 2019.

[10] A. Bhatti, B. Behinaein, D. Rodenburg, P. Hungler, and A. Etemad, "Attentive cross-modal connections for deep multimodal wearable-based emotion recognition," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, 2021, pp. 1–5.

[11] H. Cai, X. Liu, A. Jiang, R. Ni, X. Zhou, and A. Cangelosi, "Combination of EOG and EEG for emotion recognition over different window sizes," in *Proc. IEEE 2nd Int. Conf. Hum.-Mach. Syst. (ICHMS)*, Sep. 2021, pp. 1–6.

[12] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2006, pp. 1355–1358.

[13] S. Chen, L. Zhang, F. Jiang, W. Chen, J. Miao, and H. Chen, "Emotion recognition based on multiple physiological signals," *Chin. J. Med. Instrum.*, vol. 44, no. 4, pp. 283–287, 2020.

[14] M. M. Rahman et al., "Recognition of human emotions using EEG signals: A review," *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104696.

[15] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul. 2019.

[16] J. Li, Z. Zhang, and H. He, "Implementation of EEG emotion recognition system based on hierarchical convolutional neural networks," in *Proc. Int. Conf. Brain Inspired Cogn. Syst.*, 2016, pp. 22–33.

[17] P. Sarkar and A. Etemad, "Self-supervised ECG representation learning for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1541–1554, Jul. 2022.

[18] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," 2019, *arXiv:1909.11825*.

[19] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2712–2721.

[20] O. Bazgir, Z. Mohammadi, and S. A. H. Habibi, "Emotion recognition with machine learning using EEG signals," in *Proc. 25th Nat. 3rd Int. Iranian Conf. Biomed. Eng. (ICBME)*, Nov. 2018, pp. 1–5.

[21] T. Shankar, K. M. R. Kumar, and A. N. L. Jhenkar, "Analysis of EEG based emotion detection of DEAP and SEED-IV databases using SVM," *SSRN Electron. J.*, vol. 8, 2019.

[22] S. Hwang, K. Hong, G. Son, and H. Byun, "Learning CNN features from DE features for EEG-based emotion recognition," *Pattern Anal. Appl.*, vol. 23, no. 3, pp. 1323–1335, Aug. 2020.

[23] H. Chao and L. Dong, "Emotion recognition using three-dimensional feature and convolutional neural network from multichannel EEG signals," *IEEE Sensors J.*, vol. 21, no. 2, pp. 2024–2034, Jan. 2021.

[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[26] X. Jiang, J. Zhao, B. Du, and Z. Yuan, "Self-supervised contrastive learning for EEG-based sleep staging," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[27] Y. Li et al., "GMSS: Graph-based multi-task self-supervised learning for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Apr. 28, 2022, doi: 10.1109/TAFFC.2022.3170428.

[28] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 2, pp. 1–30, 2019.

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

[30] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Nov. 2013, pp. 81–84.

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

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

[33] W.-L. Zheng, H.-T. Guo, and B.-L. Lu, "Revealing critical channels and frequency bands for emotion recognition from EEG with deep belief network," in *Proc. 7th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Apr. 2015, pp. 154–157.

[34] I. Wichakam and P. Vateekul, "An evaluation of feature extraction in EEG-based emotion prediction with support vector machines," in *Proc. 11th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, May 2014, pp. 106–110.

[35] X. Jia et al., "Multi-channel EEG based emotion recognition using temporal convolutional network and broad learning system," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 2452–2457.