Motor Imagery EEG Decoding Method Based on a Discriminative Feature Learning Strategy

Lie Yang, Yonghao Song, Ke Ma, and Longhan Xie, *Member, IEEE*

***Abstract*— With the rapid development of deep learning, more and more deep learning-based motor imagery electroencephalograph (EEG) decoding methods have emerged in recent years. However, the existing deep learning-based methods usually only adopt the constraint of classification loss, which hardly obtains the features with high discrimination and limits the improvement of EEG decoding accuracy. In this paper, a discriminative feature learning strategy is proposed to improve the discrimination of features, which includes the central distance loss (CD-loss), the central vector shift strategy, and the central vector update process. First, the CD-loss is proposed to make the same class of samples converge to the corresponding central vector. Then, the central vector shift strategy extends the distance between different classes of samples in the feature space. Finally, the central vector update process is adopted to avoid the non-convergence of CD-loss and weaken the influence of the initial value of central vectors on the final results. In addition, overfitting is another severe challenge for deep learning-based EEG decoding methods. To deal with this problem, a data augmentation method based on circular translation strategy is proposed to expand the experimental datasets without introducing any extra noise or losing any information of the original data. To validate the effectiveness of the proposed method, we conduct some experiments on two public motor imagery EEG datasets (BCI competition IV 2a and 2b dataset), respectively. The comparison with current state-of-the-art methods indicates that our method achieves the highest average accuracy and good stability on the two experimental datasets.**

***Index Terms*— Motor imagery electroencephalograph (EEG) decoding, central distance loss (CD-loss), central vector shift, central vector update, circular translation strategy.**

I. Introduction

THE brain-computer interfaces (BCIs) provide a new communication approach between the human brain and external devices by decoding the electrical signals from the nervous system of brain [1]. This technology can be applied in various occasions, such as helping people who suffer from stroke, spinal cord injury, and amyotrophic lateral sclerosis, control external devices and improve their quality of life [2], [3]. Many kinds of physiological information have been applied to the BCI systems, among which motor imagery is one of the most commonly used noninvasive electroencephalograph (EEG) paradigms [4], [5]. When a person imagines or simulates physical actions, the corresponding motor imagery responses are generated in the brain with substantial neuron activity on the motor cortex [6]. Through the motor imagery based EEG decoding, disabled people can control assistive robots [7] or wheelchairs [8] to complete daily activities, such as moving and drinking, which has proved to be helpful for stroke rehabilitation [9]–[11].

There are two main processes in the motor imagery EEG decoding tasks: feature extraction and classification. Many conventional machine learning algorithms have been adopted for motor imagery EEG classification, such as random forest (RF), linear discriminant analysis (LDA), support vector machine (SVM), and so on. In research [12], Luo *et al.* proposed a motor imagery EEG decoding method by combining the dynamic frequency feature selection (DFFS) approach with the RF classifier. To overcome the singularity problem in the classical LDA, Fu *et al.* [13] proposed the regularized linear discriminant analysis (RLDA) algorithm to increase the magnitude of the diagonal elements of the scatter matrices for the motor imagery EEG decoding tasks. On the basis of classical SVM, Dong *et al.* [14] proposed a hierarchical support vector machine (HSVM) algorithm to address an EEG-based four-class motor imagery classification task. However, due to the low signal-to-noise ratio (SNR) of EEG signal, the original EEG data usually contains a lot of noise. So, the distribution of the original EEG samples is messy and the discrimination between different classes of samples is not significant enough, which is disadvantageous to the motor imagery EEG decoding tasks.

To deal with this issue, researchers have introduced a lot of feature extraction methods to extract features from motor imagery EEG samples before classification. The common spatial pattern (CSP) [15] is one of the most popular feature extraction method in the field of motor imagery EEG decoding, which can increase the difference between the extracted feature of different classes of samples. A large number of methods based on CSP have been proposed to decode motor imagery EEG accurately. For example, Novi *et al.* [16] proposed

Manuscript received October 3, 2020; revised November 25, 2020 and December 27, 2020; accepted January 12, 2021. Date of publication January 18, 2021; date of current version March 2, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 52075177, in part by the Joint Fund of the Ministry of Education for Equipment Pre-Research under Grant 6141A02033124, in part by the Research Foundation of Guangdong Province under Grant 2019A050505001 and Grant 2018KZDXM002, and in part by the Guangzhou Research Foundation under Grant 202002030324 and Grant 201903010028. *(Corresponding author: Longhan Xie.)*

Lie Yang, Yonghao Song, and Longhan Xie are with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 510460, China (e-mail: 201810100415@mail.scut.edu.cn; eeyhsong@mail.scut.edu.cn; xielonghan@gmail.com).

Ke Ma is with the State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China (e-mail: make@gzzoc.com).

Digital Object Identifier 10.1109/TNSRE.2021.3051958

sub-band CSP (SBCSP) method, which applies different band-pass filters to separate the raw EEG signal into different frequency bands and then extract features for each frequency band signal through CSP algorithm. Inspired by multiple sub-bands input idea, Ang *et al.* [17] proposed the filter bank common spatial pattern (FBCSP) that extracted the optimal spatial features through a group of band-pass filters and CSP algorithm. To improve the motor imagery EEG decoding performance of FBCSP, Thomas *et al.* [18] proposed the discriminative filter bank common spatial pattern (DFBCSP) algorithm which could obtain subject-specific discriminative filter bank instead of using fixed filter bank for all subjects. The original CSP algorithm is only suitable for binary classification, Wu *et al.*[19] presented an one-versus-rest (OVR) algorithm to extend CSP to multi-class cases. Besides the CSP-based methods, some methods based on other feature extraction algorithms also achieved excellent performance on the motor imagery EEG decoding tasks. For example, Xie *et al.* [20] proposed a simple yet efficient bilinear sub-manifold learning (BSML) algorithm to learn the intrinsic sub-manifold by identifying a bilinear mapping, and the tangent space of sub-manifold (TSSM) classification algorithm and the LDA algorithm were combined for motor imagery EEG decoding tasks. In research [21], spatio-temporal discrepancy feature (STDF) was combined with wavelet packet decomposition (WPD) for motor imagery EEG decoding tasks. Although these attempts have achieved good performance in the motor imagery EEG decoding tasks, all these methods separate feature extraction and classification into two stages. As a result, the extracted features are not the most suitable for the corresponding classifier.

In recent years, deep learning has gained extensive attention because of its excellent performance in the field of computer vision and natural language processing [22], [23]. Researchers have proposed many end-to-end motor imagery EEG decoding methods based on deep learning. For instance, Li *et al.* [24] proposed an end-to-end framework named channel-projection mixed-scale convolutional neural network (CP-MixedNet) to improve the motor imagery EEG decoding performance. In research [25], Zhao *et al.* proposed a 3D representation for motor imagery EEG data, and designed a multi-branch 3D convolutional neural network and the corresponding classification strategy for the new representation data. Moreover, an envelope representation was proposed for the motor imagery EEG data, and a convolutional neural network (CNN) was designed and optimized according to the representation for motor imagery EEG decoding in research [26]. Because of the strong ability of feature learning and embedding feature separation and classification into a single network, in general, the deep learning-based methods can achieve better performance than traditional methods. However, the existing EEG decoding methods based on deep learning only introduce the constraint of classification loss in their objective functions, so they cannot obtain the features with high discrimination, and it is difficult to further improve the accuracy of motor imagery EEG decoding.

In this paper, an end-to-end CNN framework is designed to extract both spatial and temporal features of EEG data. Inspired by the central loss of the face recognition task in reference[27], a discriminative feature learning strategy is proposed for the CNN framework to increase the discrimination of different classes of samples in the feature space. This strategy includes the central distance loss (CD-loss), the central vector shift strategy, and the central vector update process. The CD-loss is proposed to promote the same class of samples to gather around the corresponding central vector in the feature space. And the central vector shift strategy is proposed to increase the distance between different classes of samples in the feature space, which can significantly improve the discrimination of different classes of samples in the feature space. In addition, to avoid the non-convergence of CD-loss and weaken the influence of the initial value of the central vector on the final result, the central vector update process is introduced in the proposed framework. Different from the traditional EEG decoding methods, the feature extraction part and the classification part of our method are optimized according to the same objective function, which is conducive to obtain more suitable features for classification and obtain higher EEG decoding accuracy. Unlike other motor imagery EEG decoding methods based on deep learning, we creatively propose the CD-loss, the central vector shift strategy, and the central vector update process, which are helpful to obtain more discriminative features and achieve better classification performance.

In addition, overfitting is another severe challenge for motor imagery EEG decoding methods based on deep learning. To deal with this problem, researchers have proposed some data augmentation methods for EEG data. For example, in reference [24], the Gaussian noise with the mean value of 0 and the standard deviation of 0.001 was added to the original EEG samples for data augmentation. Although this kind of methods could alleviate the overfitting problem to a certain extent, these methods would introduce some redundant noise, which was harmful to stability of the motor imagery EEG decoding process. Some researchers also took part of the data from the original samples as new samples. For example, Guennec *et al.* [28] extracted slices from the original samples in the time series and perform EEG decoding at the slice level. This kind of methods could extend the training set to alleviate the overfitting problem. However, each sample only contains partial data of the original sample, which was not conducive to the improvement of EEG decoding accuracy. Considering the shortcomings of these methods, a novel data augmentation method based on circular translation is proposed in this paper. The proposed data augmentation method can greatly expand the dataset without introducing any additional noise and each sample contains all the data of the corresponding original sample. So, this data augmentation method can significantly alleviate overfitting and improve the generalization ability of our motor imagery EEG decoding model.

In summary, the main contributions of this work are summarized as follows: first, a discriminative feature learning strategy is proposed for motor imagery EEG decoding tasks.

This strategy can improve the discrimination of different classes of samples in the feature space and improve the EEG decoding accuracy to a large extent. Furthermore, to alleviate

the overfitting problem, a data augmentation method based on the circular translation strategy is proposed to expand the original dataset without losing any information of the original samples or introducing any extra noise. In addition, we present a CNN framework based on the discriminative feature learning strategy and the proposed data augmentation method for end-to-end motor imagery EEG decoding.

II. Methodology

In this part, the proposed data augmentation method based on the circular translation strategy is firstly introduced. Then we describe the working principle of the proposed discriminative feature learning strategy in detail. Next, we present the specific network architecture of the motor imagery EEG decoding method in this paper. And the training process of the motor imagery EEG decoding framework based on the discriminative feature learning strategy is introduced finally.

A. Data Augmentation With Circular Translation Strategy

In this paper, each sample of EEG data based on motor imagery is represented as a 2D matrix of $C \times T$, in which the rows represent the data collected from different electrodes and the columns are the data at different sampling time points (as shown in Fig. 1 (a)).

Since the amount of EEG data is very limited, overfitting is one of the most important problems encountered during the network training process. In this paper, a novel data augmentation method based on the circular translation strategy is proposed to alleviate this problem. During the data augmentation process, we move the samples circularly in the time dimension with a step size of D, while the arrangement of electrodes remaining unchanged. In the time dimension, the initial sample is $0 \sim T$, the first circular translation sample is spliced by D $\sim T$ and $0 \sim D$, the second circular translation sample is spliced by $2D \sim T$ and $0 \sim 2D$, ……, and the k-th circular translation sample is spliced by $kD \sim T$ and $0 \sim kD$ (as shown in Fig. 1 (b)). It can be seen that the new samples obtained by circular translation have the same size as the original samples, and the new samples generated by the same sample only have some staggered positions in the time dimension. Therefore, the obtained samples retain the temporal and spatial features of the original samples and there are some differences between the obtained samples and the corresponding original samples. After data augmentation, the obtained samples are directly fed into the network without any other preprocessing steps.

B. Working Principle of the Proposed Discriminative Feature Learning Strategy

At present, the EEG decoding methods based on deep learning are often optimized only by the classification loss to find a hyperplane to divide input samples into different classes, but ignore the process of feature extraction. However, in the feature space, the distribution of the extracted features by these methods is always scattered, which is disadvantageous for the classification tasks. To address this problem, the CD-loss is designed to centralize the distribution of features extracted from the same class of samples to improve the classification accuracy. The CD-loss is the average distance between the feature vectors of a batch of samples and their corresponding central vectors, which is calculated as follow:

$$L_{cen} = \frac{1}{n_b}\sum_{i=1}^{n_b} ||f_i^{\mathrm{t}} - cen_{y_i}^t||_2 \tag{1}$$

where f_i^t represents the feature vector of the i-th sample in the t-th iteration; y_i represents the class of the i-th sample in the training batch; $cen_{y_i}^t$ represents the central vector of the y_i-th class of samples during the t-th iteration; and n_b denotes the number of samples in a batch.

Before the training process, we need to initialize the central vectors of each class of samples at first. In this paper, the average value of the initial feature vectors extracted from each class of samples in the training set are taken as the initial values of the corresponding central vectors (as shown in Fig. 2(a)). The expression of the central vector initialization process is:

$$cen_j^0 = \frac{\sum_{i=1}^{n_s}\delta(y_i = j)\cdot f_i^0}{1 + \sum_{i=1}^{n_s}\delta(y_i = j)} \tag{2}$$

where cen_j^0 is the initial central vector corresponding to the samples with the label of j; n_s denotes the total number of samples in the training set; $\delta(y_i = j) = \begin{cases} 0, if\, y_i \neq j \\ 1, if\, y_i = j \end{cases}$; y_i represents the class of the i-th sample in the training set; and f_i^0 is the initial feature vector of the i-th sample in the training set.

Through the constraint of CD-loss, the distance between feature vectors of the same class of samples is gradually reduced, and each class of samples is eventually clustered near the corresponding central vector (as shown in Fig. 2 (b)). However, the samples whose feature vectors are relatively far away from the corresponding central vector are still easy to be misclassified because the distance between different central vectors is not large enough. To deal with this issue, the central vector shift strategy is proposed to increase the distance between the central vectors of different classes. First, we calculate the average vector c^t of all the central vectors during the t-th iteration and then shift each central vector cen_i^t along the direction of $\overrightarrow{c^t cen_i^t}$ by a certain step size (as shown in Fig. 2 (c)). The expression of the central vector shift process is as follows:

$$c^t = \frac{1}{n_c}\sum_{i=1}^{n_c} cen_i^t \tag{3}$$

$$cen_i^{t+1} = cen_i^t + \alpha\cdot\frac{(cen_{y_i}^t - c^t)}{||cen_{y_i}^t - c^t||_2} \tag{4}$$

where c^t is the average vector of all central vectors during the t-th iteration; cen_i^t and cen_i^{t+1} are the central vectors of samples with the label of i before and after the central vector shift process in the t-th iteration, respectively; n_c is the number of classes of all samples in the experimental datasets; and α is the step size of the central vector shift process.

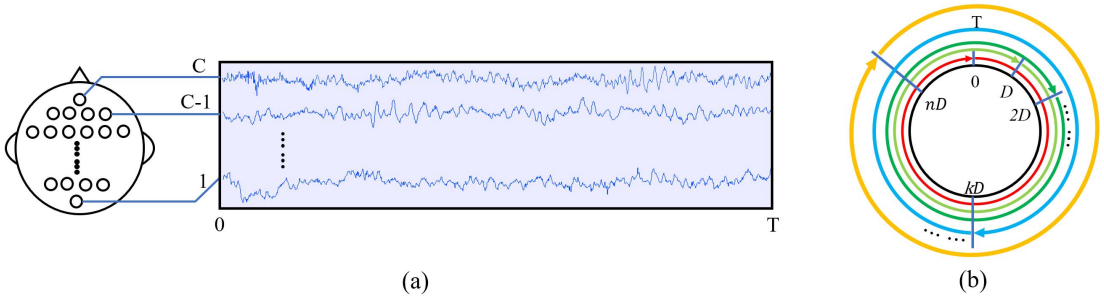

Fig. 1. The representation of each EEG sample and the data augmentation method based on the circular translation strategy. (a) is the data representation in this paper, C denotes the number of channels per sample, T is the number of sampling points contained in each sample; (b) is the data augmentation method based on the circular translation strategy.

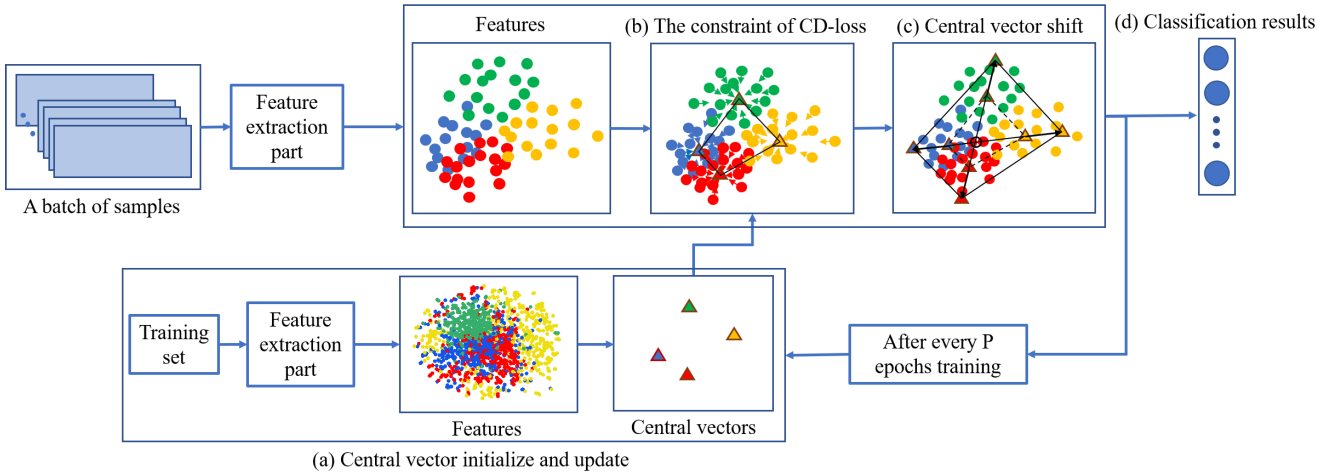

Fig. 2. The working principle schematic diagram of the proposed EEG decoding method. The solid dots with different colors represent the feature vectors of different classes of samples. (a) The initialization and update process of the central vectors; (b) The result with the constraint of the CD-loss; (c) The central vector shift process; (d) The classification results.

During the training process, if the speed of the central vector shift process is faster than that of the feature vectors converging to the corresponding central vectors, the CD-loss will be increasingly larger and become nonconvergent. To avoid this situation, the central vector update process is proposed in this paper. After P epochs training, the corresponding central vectors are updated by the average values of the feature vectors extracted from each class of samples (as shown in Fig. 2 (a)). The expression of the central vector update process is:

$$cen_j^{0,e} = \frac{\sum_{i=1}^{n_s} \delta(y_i = j) \cdot f_i^{0,e}}{1 + \sum_{i=1}^{n_s} \delta(y_i = j)} \tag{5}$$

where $cen_j^{0,e}$ represents the initial central vector corresponding to the samples of the j-th class in the e-th epoch and $f_i^{0,e}$ denotes the initial feature vector of the i-th sample in the e-th epoch. The update of the central vectors can not only avoid the nonconvergence of the CD-loss caused by the fast shift speed of the central vectors but also weaken the influence of the initial value of the central vectors on the final classification results. Through the proposed discriminative feature learning strategy, we can obtain the features with very high discrimination. Then classification is conducted according to these discriminative features to realize motor imagery EEG decoding with high accuracy (as shown in Fig. 2 (d)).

C. Network Architecture

The representation of EEG data in this paper is the same as that of the single-channel image data represented as a 2D digital matrix but with different meanings for the rows and columns. For the EEG data, each row represents the data collected from different electrodes, and each column represents data of different sampling time, while the rows and columns of the image represent the locations of pixels. Therefore, some typical CNN frameworks that are applied for the classification tasks in the field of computer vision, such as LeNet [29], AlexNet [30], VGG [31], or ResNet [32], cannot directly perform well on the original EEG data.

In this paper, a CNN framework is proposed for the motor imagery EEG decoding tasks based on the discriminative feature learning strategy and the proposed data augmentation method, which mainly includes a convolutional part and a fully connected part (as shown in Fig. 3). The convolutional part is composed of a temporal convolution module, a spatial convolution module and a general convolution module. The temporal convolution module with the convolutional kernel

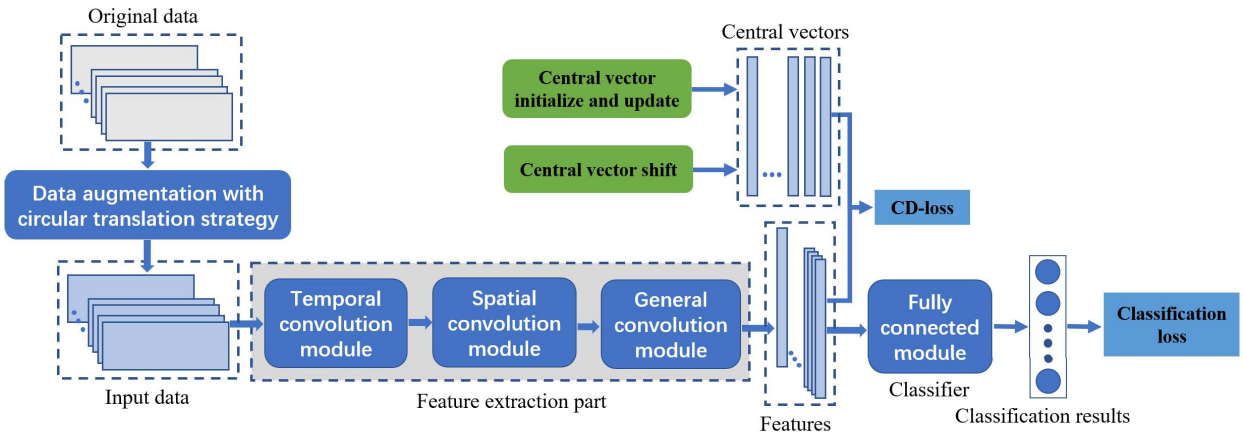

Fig. 3. The overall network structure of our motor imagery EEG decoding method.

size of $1 \times m$ is adopted to extract the temporal features, and the spatial convolutional module with the convolutional kernel size of $n \times 1$ is applied for the spatial feature extraction. After these two modules, a general convolution module is introduced for feature integration and high-level features extraction. Following this module, a fully connected module is adopted for classification according to the features extracted by the convolutional part.

The specific network structure of the proposed EEG decoding method is shown in Table I. The temporal convolutional module contains a convolutional layer with a kernel size of 1×23, and the stride step size of the convolutional kernel is (1, 1). The main function of this module is to extract the temporal features from the input EEG samples. The spatial convolutional module includes a spatial convolutional layer with the kernel size and stride step size of $C \times 1$ and (1, 1) respectively, where C is the number of channels for the input EEG data. This module is introduced to extract the spatial features of different channels. A general convolutional module with two convolutional layers and two max-pooling layers is added following these two modules to improve the learning ability of the framework and integrate the extracted temporal and spatial features. The first convolutional layer with the kernel size of 1×17 and the stride step size of (1, 1) is adopted to extract features with a larger scale. The second convolutional layer with the kernel size of 1×7 and the stride step size of (1, 1) is used for small scale features extraction. Each convolutional layer of this module is followed by a max-pooling layer with the kernel size and stride step size of 1×6 and (1,6), respectively. There is only one fully connected layer in the fully connected module, which is introduced for the classification of input samples according to the features extracted by the convolutional part.

D. Training Process of the Proposed CNN Framework

The training process of the proposed EEG decoding framework is described as Algorithm 1. Before the training process, the network parameters are initialized according to standard normal distribution, and the central vector is initialized according to equation (2). Then, the classification loss is adopted to promote features of samples in the same class to converge into the same region, and the proposed CD-loss is adopted to improve the feature discrimination of different classes of samples (as show in Fig. 3). In each iteration of the training process, the central vectors shift process is conducted according to equation (3) and (4) to improve the performance and robustness of the proposed framework. In addition, the central vectors are updated according to equation (5) to avoid the non-convergence of CD-loss and weaken the influence of the central vector initial value after every P epochs of training.

III. Experiments

A. Experimental Dataset

1) BCI Competition IV 2a Dataset: The BCI competition IV 2a dataset [33], provided by Graz University, is used to evaluate the effectiveness of our motor imagery EEG decoding method. The dataset contains EEG data from 9 healthy subjects performing 4 different motor imagery tasks: movement of the left hand, right hand, both feet and tongue. The signals were recorded by 22 EEG electrodes at a 250 Hz sampling frequency and then bandpass filtered between 0.5 Hz and 100 Hz and notch filtered at 50 Hz. 2 sessions on different days were recorded for each subject, and each session comprised 288 trials. The sampling period of each trial is 3s, which results in 750 sample points for each trial. In this paper, we take each trial as a sample, and each sample is represented as a 2D-matrix of 22×750. The 22 rows of each sample represent signals recorded from the 22 electrodes, and the 750 columns of a sample represent the EEG data of the 750 sample points.

2) BCI Competition IV 2b Dataset: The BCI competition IV 2b dataset [34] includes two classes (motor imagery of left hand and right hand) EEG data from 9 subjects of a study published in [35]. For each subject, there are 5 sessions, in which the first 2 sessions contain data without feedback and the last 3 sessions were recorded with feedback. Each session without feedback consisted of 6 runs, and each run includes 10 trials of each kind of motor imagery task. This resulted in 120 trials of each subject per session. The subjects had to imagine the corresponding hand movement over a period of 3s. Each trial was followed by a short break of at least 1.5s. For

TABLE I
THE SPECIFIC NETWORK STRUCTURE OF THE PROPOSED EEG DECODING FRAMEWORK. *C* IS THE NUMBER OF CHANNELS FOR THE INPUT EEG DATA

Module	Temporal convolution module	Spatial convolution module	General convolution module				Fully connected module
Layer	Conv layer	Conv layer	Conv layer	MP layer	Conv layer	MP layer	FC layer
Kernel	1×23	C ×1	1×17	1×6	1×7	1×6	---
Stride	(1, 1)	(1, 1)	(1, 1)	(1, 6)	(1, 1)	(1, 6)	---
Feature maps	10	30	30	30	30	30	---

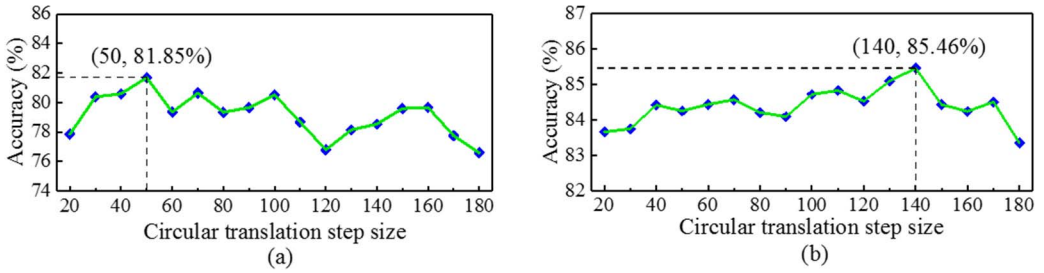

Fig. 4. The experimental results of data augmentation with different circular translation step sizes. (a) is the experimental result of the BCI competition IV 2a dataset; (b) is the experimental result of the BCI competition IV 2b dataset.

TABLE II
THE EXPERIMENTAL RESULTS OF THE BASELINE, EXPERIMENT 1 ~ 4 ON THE BCI COMPETITION IV 2A DATASET

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	Average accuracy	Standard deviation
Baseline	84.44	62.73	90	72.22	64.34	58.89	87.78	85.86	69.6	75.1	11.33
Experiment 1	91.82	71.41	91.52	73.64	78.48	59.9	95.45	91.31	77.88	81.26	11.3
Experiment 2	91.72	72.73	93.43	75.15	78.89	58.79	95.05	90.91	75.96	81.4	11.49
Experiment 3	93.23	71.82	92.32	77.58	79.6	53.33	94.34	89.49	78.89	81.18	12.45
Experiment 4	91.31	71.62	92.32	78.38	80.1	61.62	92.63	90.3	78.38	81.85	10.15

the 3 online feedback sessions, 4 runs with smiley feedback were recorded, whereby each run consisted of 20 trials for each type of motor imagery. There were 160 trials of every subject in each of the last three sessions and the feedback period of each trial last 4s. All the EEG data in this dataset were recorded from 3 channels (channel C3, Cz and C4) with the sampling frequency of 250 Hz. The data from 4s to 7s of each trial is intercepted as a sample in our works, which results in 750 sampling points of each sample. As described in Section II.A, each sample is represented as a 2D matrix of 3×750.

B. Data Augmentation

To alleviate the overfitting problem, we expand the experimental datasets through the circular translation strategy described in Section II.A. In the data augmentation process, the step size of circular translation is a very important parameter. If the step size is too large, we cannot obtain enough augmented data to overcome the overfitting problem. However, if the step size is too small, the difference between different samples becomes very small, which is disadvantageous to improve the generalization of the proposed framework. To select a suitable step size for each dataset, we conduct some experiments on the two experimental datasets with many different step sizes. According to the experimental results (as shown in Fig. 4), the circular translation step size of BCI competition IV 2a and 2b dataset are select as 50 and 140, respectively.

1) Experimental Implementation Details: Similar to [25], the original data of each subject in the BCI competition IV 2a dataset is randomly divided into 10 subsets of equal size. Then, 10-fold cross-validation is conducted for each subject. In the experiments of the BCI competition IV 2b dataset, the same as [36], we take the 400 trails of session 1~3 for each subject as the training set, and the 320 trails of session 4 and 5 for each subject is taken as the test set. Before the training process, the training set of each subject is augmented according to the data augmentation method based on the circular translation strategy.

To further alleviate overfitting, dropout layer with the probability of 0.38 is added following each convolutional layer of the proposed framework. The activation functions of all the convolutional layers are selected as ELUs. In addition, the Adam optimizer is selected to optimize the proposed framework, and the parameters of the optimizer are set as $\beta_1 = 0.5$, $\beta_2 = 0.999$. During the training process, the learning rate is set to 0.0005, and the batch size is set to 24. Moreover, the central vector shift is conducted every epoch with the shift

TABLE III
THE ACCURACY COMPARISON OF THE PROPOSED METHOD AND THE CURRENT STATE-OF-THE-ART METHODS ON BCI COMPETITION IV 2A DATASETS. S 1∼9 DENOTE THE 9 SUBJECTS IN THE DATASET, RESPECTIVELY

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	Average accuracy	Standard deviation
Multi-Branch 3D CNN [25]	77.4	60.14	82.93	72.29	75.84	68.99	76.04	76.85	84.66	75.015	6.92
TSSM+LDA [20]	81.8	62.5	88.8	63.7	62.9	58.5	86.6	85.1	90.0	75.5	12.46
Envelope + CNN [26]	85.23	69.73	90.15	65.57	77.42	52.41	93.68	90.04	84.75	78.78	12.95
Our method	**91.31**	**71.62**	**92.32**	**78.38**	**80.1**	**61.62**	**92.63**	**90.3**	**78.38**	**81.85**	**10.15**

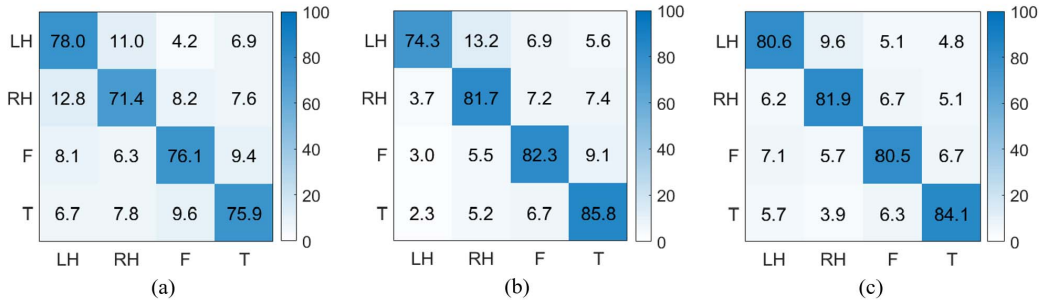

Fig. 5. The confusion matrices of the Baseline, Experiment 1, and Experiment 4 on the BCI competition IV 2a dataset. In each confusion matrix, *LH*, *RH*, *F*, and *T* denote the movement imagery tasks of left hand, right hand, both feet and tongue, respectively. (a) is the confusion matrix of Baseline, (b) is the confusion matrix of Experiment 1, and (c) is the confusion matrix of Experiment 4.

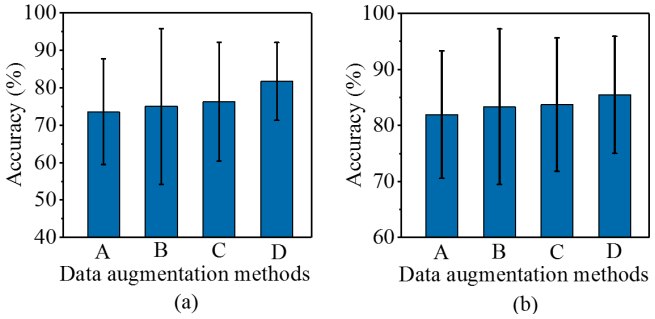

Fig. 6. The experimental results of different data augmentation methods. A is the result of experiment without data augmentation; B is the experimental result of data augmentation with Gaussian noise; C is the experimental result of data augmentation with sliding windows; D is the experimental results of the data augmentation method proposed in this paper. (a) are the experimental results on the BCI competition IV 2a dataset; (b) are the experimental results on the BCI competition IV 2b dataset.

step size of 0.002, the period of central vector updating process is 20 epochs, and the weight λ of central loss in the full objective function is set as $\lambda = 10$ in this paper. In order to speed up the training process, the proposed CNN framework is implemented with *PyTorch* on the GeForce 2080ti platform.

IV. EXPERIMENTAL RESULTS

A. Evaluation of the Proposed Discriminative Feature Learning Strategy

To evaluate the effectiveness of the CD-loss, the central vector shift strategy, and the central vector update process of the proposed discriminative feature learning strategy, we carry out several comparative experiments on the BCI competition IV 2a dataset in this section. The experimental results are shown in Table II. First, the proposed framework with only the constraints of classification loss is taken as the Baseline, of which the result is $75.1 \pm 11.33\%$. Then, we adopt the constraint of CD-loss to the Baseline and take it as Experiment 1. According to the table, the average accuracy of Experiment 1 is 6.16% higher than that of Baseline, which proves that the proposed CD-loss is very effective to improve the performance of the motor imagery EEG decoding tasks. Next, we introduce the central vector shift strategy on the basis of Experiment 1 and take it as Experiment 2. It can be seen from the experimental results of Experiment 1 and Experiment 2 that by introducing the central vector shift strategy individually, although the average accuracy is slightly improved, the standard deviation of accuracy for all subjects increases a little as well. And the experiment by introducing the central vector update process on the basis of Experiment 1 is taken as Experiment 3. According to the experimental results of Experiment 1 and Experiment 3, we know that both the average accuracy and the standard deviation of accuracy are reduced by introducing the central vector update process separately. Finally, the central vector shift strategy and central vector updating strategy are adopted to Experiment 1 together, which is taken as Experiment 4. According the results comparison of Experiment $1 \sim 4$, it can be concluded that introducing the central vector shift strategy and the central vector update process separately is difficult to improve the experimental results. And the performance and stability of our method can be improved by introducing the central vector shift strategy together with the central vector update process.

In addition, we present the confusion matrices of the Baseline, Experiment 1, and Experiment 4 on the BCI competition IV 2a dataset, respectively (as shown in Fig. 5). The comparison of Fig. 5 (a) and (b) indicates that CD-loss

TABLE IV
THE ACCURACY COMPARISON OF THE PROPOSED METHOD AND THE CURRENT STATE-OF-THE-ART METHODS ON BCI COMPETITION IV 2B DATASETS. S 1~9 DENOTE THE 9 SUBJECTS IN THE DATASET, RESPECTIVELY

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	Average accuracy	Standard deviation
RSMM [35]	72.5	56.43	55.63	97.19	88.44	78.75	77.5	91.88	83.44	77.97	13.73
FDBN [36]	81	65	66	98	93	88	82	94	91	84	11.25
RF with DFFS [12]	73.24	67.48	63.01	97.4	95.49	86.66	84.68	95.93	92.61	84.06	12.31
WPD + STDF [21]	69.5	64	86.5	96	94	87	83	95.5	92	85.28	10.81
Our method	**79.69**	**60.71**	**82.19**	**96.87**	**94.37**	**89.37**	**82.19**	**93.75**	**90**	**85.46**	**10.44**

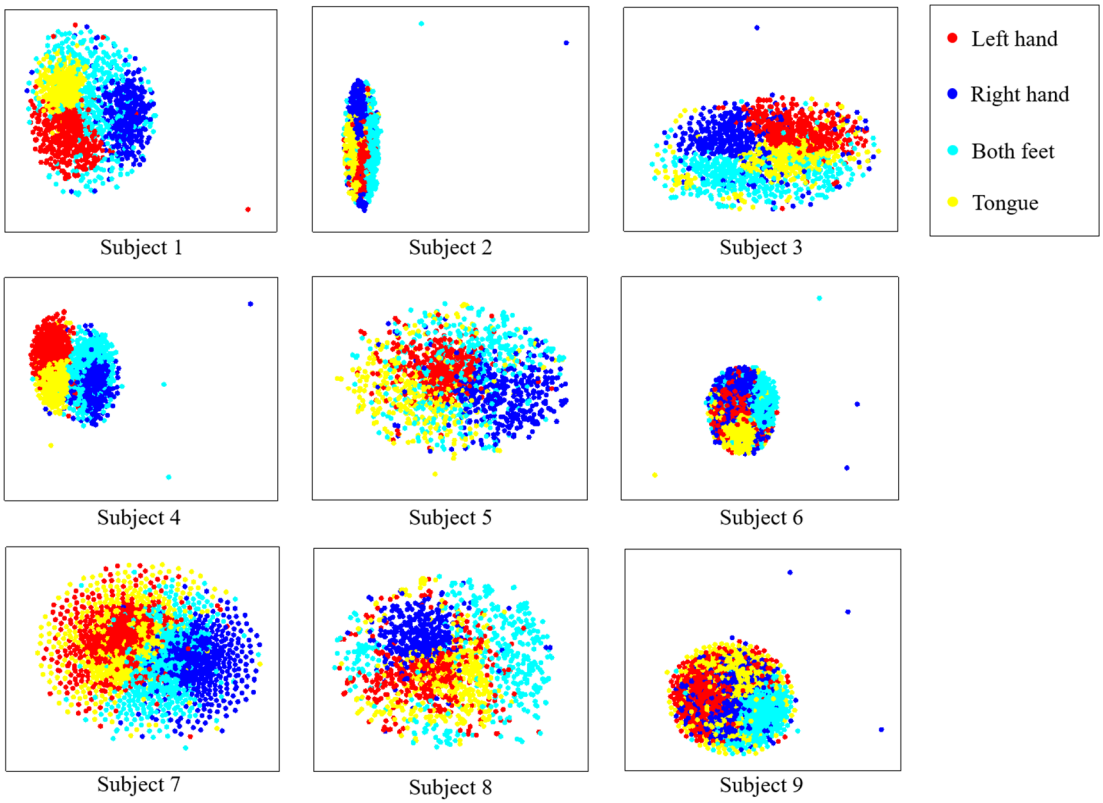

Fig. 7. The features that are obtained by the feature extraction part of the proposed EEG decoding framework in Baseline, mapped to 2D plane by TSNE.

plays a significant role in accuracy improvement of the EEG decoding tasks. The confusion matrix in Fig. 5 (c) shows that, by adopting the central vector shift strategy and central vector update process, the average accuracy is improved, and the accuracy gap between different classes has narrowed. These experimental results further prove that the proposed CD-loss can greatly improve the EEG decoding accuracy. Moreover, by adopting the central vector shift strategy and central vector update process, the experimental results can be further improved and the accuracy of every class will become more balanced.

B. Evaluation of Our Data Augmentations Method

To evaluate the data augmentation method proposed in this paper, we conduct some experiments with different data augmentation methods on the BCI competition IV 2a and 2b dataset, respectively. First, we conduct the experiment without data augmentation as the control group. Then, under the same conditions, experiments are carried out on the data obtained through the data augmentation method by adding Gaussian noise, the data augmentation method based on window sliding and the proposed data augmentation method based on circular translation strategy. As shown in Fig. 6, A is the experimental result of the control group. B is the result of experiment with the data augmentation method by adding Gaussian noise, of which the standard deviation is set to 0.001 according to [24]. C is the result of the data augmentation method based on window sliding, and in this paper, the length of each windows and the sliding step size are set to 600 and 10, respectively. D is the experimental results of our data augmentation method. As illustrated in Fig. 6, the proposed data augmentation method achieves higher average accuracy and lower accuracy standard deviation than those of other data augmentation methods, which demonstrates that the proposed data augmentation method can alleviate overfitting to a large extent and helps achieve better and more stable performance than other methods.

C. Comparison With the-State-of-the-Art Methods

In this section, we conduct some experiments on BCI competition IV 2a and 2b dataset respectively, and compare the experimental results with that of the current state-of-the-art

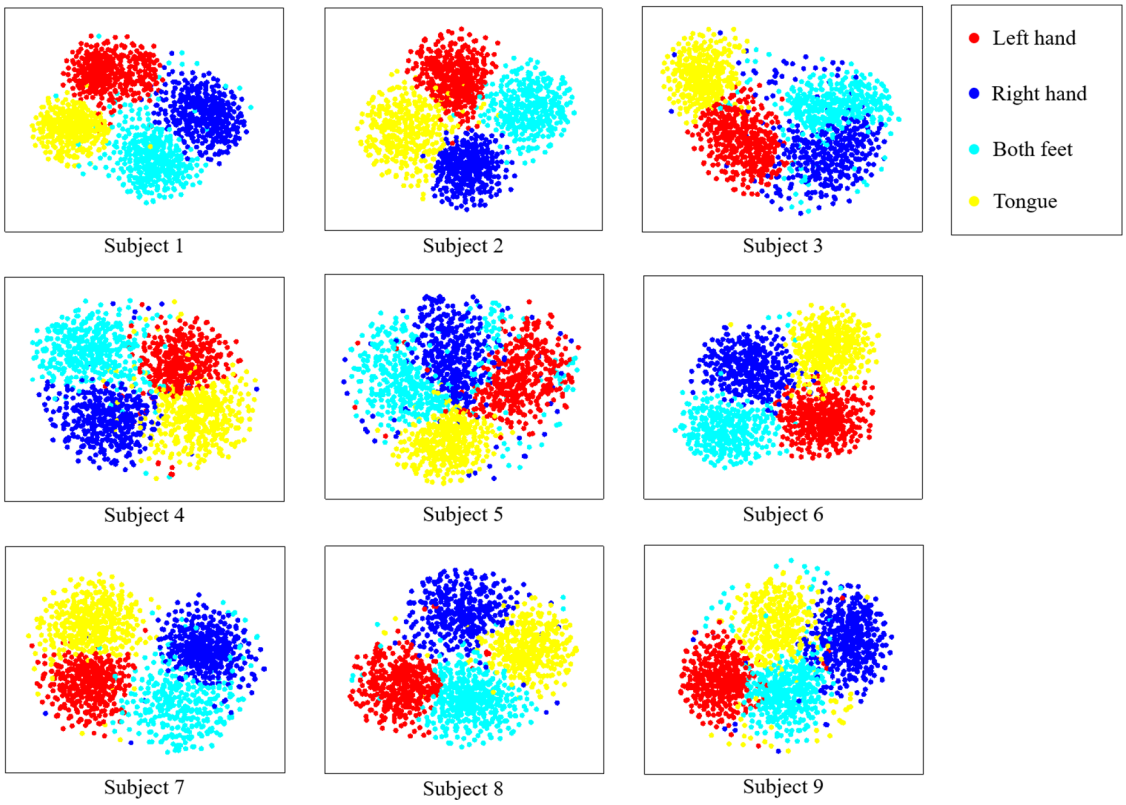

Fig. 8. The features that are obtained by the feature extraction part of the proposed EEG decoding framework in Experiment 1, mapped to 2D plane by TSNE.

methods to further prove the effectiveness of the proposed EEG decoding method. It is well known that there are huge individual differences between EEG signals of different subjects. To overcome the individual difference, the same as many researches such as [21], [22], and [25], we train a model for each subject separately.

The motor imagery EEG decoding accuracy of each subjects and their average accuracy on this dataset are shown in Table III. In this table, we evaluate the proposed algorithms against the competing algorithms on the BCI competition IV 2a dataset, including Multi-Branch 3D CNN [25], TSSM + LDA [20], and Envelope + CNN [26]. As we can see in this table that our method can achieve higher accuracy than all the competing methods on the majority of subjects except for S6, S7, and S9. Moreover, the proposed EEG decoding method achieves the highest average accuracy on the BCI competition IV 2a dataset. Although the accuracy standard deviation of multi-branch 3D CNN is 3.23 lower than that of our method, its average accuracy is 6.835% lower than that of our method. And the accuracy standard deviation of our method is lower than that of other competing methods. In general, our method has the best performance and very good stability on the BCI competition IV 2a dataset.

As shown in Table IV, the experimental results on the BCI competition IV 2b dataset of our method is compared with that of some state-of-the-art methods, such as RSMM [36], FDBN [37], RF with DFFS [12], and WPD + STDF [21]. Because of the huge difference between EEG signals of different subjects, many EEG decoding methods are not stable for different subjects, which achieve very high accuracy on some subjects, but achieve low accuracy on other subjects. It can be seen from Table IV that some methods achieve higher classification accuracy than our method on some subjects. However, our method has higher average accuracy and lower standard deviation than all the competing state-of-the-art methods on this dataset. These results prove that our method performs better than all the competing state-of-the-art methods, and it is more robust to different experimental subjects than other methods on the BCI competition IV 2b dataset.

V. Discussion

At present, a number of end-to-end EEG decoding methods based on deep learning have emerged. However, almost all these methods adopt only the classification loss in their objective functions. As can be seen from Fig. 7, only under the constraint of classification loss, the distribution of the obtained features in the feature space is chaotic, and the features of different classes of samples is not discriminative enough, so it is difficult to achieve high classification accuracy. In order to increase the discrimination of different classes of samples in the feature space, the discriminative feature learning strategy is proposed in this paper, which includes the CD-loss, the central vector shift strategy, and the central vector update process.

The discriminative feature learning strategy is inspired by the central loss of the face recognition task in reference[27], but it is obviously different from the central loss. Firstly, the central loss can only make samples of the same class converge to the same region, but cannot increase the distance between different classes of samples in the feature space. Due to the low discrimination of different classes of motor imagery

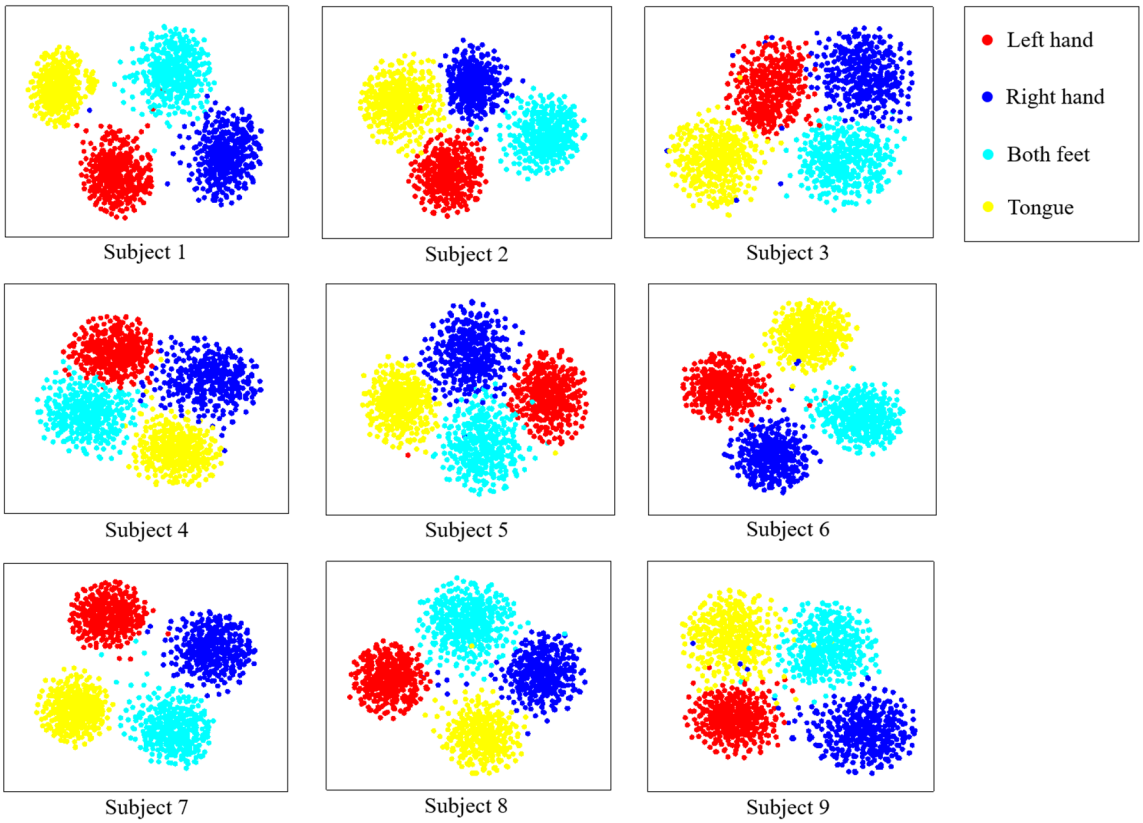

Fig. 9. The features that are obtained by the feature extraction part of the proposed EEG decoding framework in Experiment 4, mapped to 2D plane by TSNE.

EEG samples, the distribution regions of different classes of samples in the feature space often have some intersections, which is very disadvantageous to the decoding task. In order to improve the discrimination of different classes of samples in the feature space, the central vector shift process is introduced in our discriminative feature learning strategy. In addition, the central vectors of central loss are updated according to each batch of samples. Due to the low SNR of the motor imagery EEG data, this update process will lead to confusion in the update direction of the central vectors, which makes the central loss continuously oscillate and difficult to converge. To ensure that the convergence process of CD-loss is more stable, our discriminative feature learning strategy updates the central vectors according to the whole training set.

With the constraint of CD-loss, the features of each sample converge to the central vector of the corresponding class. As shown in Fig. 8, features of the same class of samples will eventually gather to the same region of the feature space, and the features of samples in different classes are distributed in different regions, which indicates that the discrimination of features extracted from different classes of sample is significantly enhanced. Therefore, according to the results of Baseline and Experiment 1 in Table II, the average accuracy is improved from 75.1% to 81.26% after introducing the CD-loss, which confirms that the proposed CD-loss is very effective in improving the performance of EEG decoding tasks.

In order to increase the distance between the feature distribution regions of different classes of samples and further improve the discrimination of different classes of samples in the feature space, the central vector shift strategy is proposed in our works. In addition, we update the central vectors according to the obtained features of all samples after every 20 epochs training to prevent the CD-loss from non-convergence due to the fast speed of the central vector shift. According to the comparison of Fig. 8 and Fig. 9, the distance between the regions of different classes of samples increases significantly after introducing the central vector shift strategy together with the central vector update process. Moreover, the experimental results of Experiment 1 and 4 in Table II show that the central vector shift strategy and the central vector update process not only improve the average accuracy, but also reduce the accuracy standard deviation of different subjects. The results demonstrate that the central vector shift strategy together with the central vector update process can promote our method to achieve better and more robust performance.

To overcome the overfitting problem, a data augmentation method based on the circular translation strategy is proposed in this paper, which neither introduces additional noise as the method based on adding random noise [24] nor loses part of the information as the method based on sliding window [28]. According to the experimental results in Fig. 6, by comparing the experimental results B with A, we know that the data augmentation method by adding Gaussian noise is helpful to improve the average accuracy, but the accuracy standard deviation of this method is significantly higher than that of other methods. This is because the data augmentation method will introduce redundant noise, which leads to the poor stability of the EEG decoding process. From the experimental results

comparison of A, C, and D, we can see that the average accuracy is improved by adopting the data augmentation method based on window sliding. But each sample obtained by this method contains only partial data of the corresponding original sample, which limits the improvement of decoding accuracy. The proposed data augmentation method based on the circular translation strategy can obtain a large number of samples that have the same size as the original samples, without any data loss. Therefore, our data augmentation method significantly improves the accuracy and stability of the EEG decoding process.

In addition, the experimental results comparison with the state-of-the-art methods in Table III and Table IV proves that the proposed data augmentation method alleviates the overfitting problem to a large extent, and the discriminative feature learning strategy promotes our CNN framework to achieve better performance and good stability on the motor imagery EEG decoding tasks.

However, the proposed motor imagery EEG decoding method still suffers from some limitations. First, in the central vector update process, we need to calculate the mean vector of all the feature vectors extracted from the training samples, which consumes many computing resources. In addition, each input sample of our motor imagery EEG decoding method contains the EEG signal lasting for 3s in the time dimension, which will lead to the delay of our method in the online system. Therefore, in the future work, we will optimize the discriminative feature learning strategy to reduce the amount of computation of our method. And we will further reduce the length of the input sample in the time dimension while maintaining high motor imagery EEG decoding accuracy.

VI. Conclusion

In summary, we propose a motor imagery EEG decoding method based on the discriminative feature learning strategy and the circular translation data augmentation method in this paper. First, the discriminative feature learning strategy is proposed for the motor imagery EEG decoding network to increase the discrimination of different classes of samples in the feature space, which helps improve the decoding accuracy to a large extent. Then, a data augmentation method based on circular translation strategy is proposed to alleviate the overfitting problem. The experimental results on the two public datasets (the BCI competition IV 2a and 2b dataset) show that our method achieves better performance than the compared state-of-the-art methods, and has good stability for different objects. These results confirm that the proposed method can be regarded as a potential approach to improve the performance of motor imagery EEG-based BCI systems.

Appendix

Evaluate the Proposed Discriminative Feature Learning Strategy at the Feature Level

In order to further explore the role of CD-loss at the feature level, we map the features that are obtained by the feature extraction part of the proposed framework in Experiment 1 and Baseline to a 2D plane by TSNE [38] for every subject (as shown in Fig. 7 and Fig. 8). It can be seen from Fig. 7 that when CD-loss is not introduced, the distribution of the obtained features is rather chaotic, and the features of samples in different classes are interwoven in the feature space. However, as we can see in Fig. 8 that with the constraint of CD-loss, features of the same class of samples converge to the same region in the feature space, and the features of different class of samples become more distinguishable, which is very beneficial to the classification tasks.

In addition, to further study the effect of the central vector shift strategy together with the central vector update process on the EEG decoding tasks, the features of each subject obtained from the feature extraction part in Experiment 4 are mapped to a 2D plane through TSNE [38] (as shown in the Fig. 9). The comparison of Fig. 8 and Fig. 9 indicates that the distance between features of sample in different classes increases significantly by introducing the central vector shift strategy and the central vector update process. The results show that the central vector shift strategy together with the central vector

Algorithm 1 Optimization Algorithm of the Proposed Framework

Input: Training set $D = \{D_i\}_{i=1}^{s}$; the batch size B, the learning rate of generator and discriminator optimization lr, the type of optimizer Adam, the parameters of the optimizer β_1, β_2; The number of classes n_c; the weight of CD-loss in the full objective function λ.

Output: Network parameters of the proposed CNN framework W.

Initialize: Network parameters W are initialized according to standard normal distribution; the central vectors $cen_i (i = 1, 2, 3, \ldots, n_c)$ are initialized according to equation (2).

Repeat:

1. Update the central vectors after every P epochs of training according to equation (5).
2. **for** $t = 1, 2, \ldots, s$ **do**
3. $D_t = \{x, y,$ x is a batch of sample, y is the corresponding labels of x.
4. Extract features of input data x with the feature extraction part $Conv$: $Conv(x) \rightarrow Features$, and classification through the fully connected part Cls: $Cls(Features) \rightarrow c_p$.
5. Calculating the classification loss is: $L_{cls} = E_{x,y}[-logCls(y \mid Conv(x))]$.
6. Calculating the CD-loss is: $L_{cen} = \frac{1}{B}\sum_{i=1}^{B} ||f_i - cen_{y_i}||_2$, where $\{f_i\}_{i=1}^{B} = features and cen_k$ is the central vector of the k-th class.
7. Optimize the network parameters of the proposed CNN framework W according to the full objective function $Loss$ through the optimizer: $Loss = L_{cls} + \lambda \cdot L_{cen}$.
8. Central vector shift according to equation (3) and (4).
9. **end for**

Until convergence

update process further increases the discrimination of samples in different classes.

REFERENCES

[1] L. He, D. Hu, M. Wan, Y. Wen, K. M. von Deneen, and M. Zhou, "Common Bayesian network for classification of EEG-based multiclass motor imagery BCI," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 46, no. 6, pp. 843–854, Jun. 2016, doi: 10.1109/TSMC.2015.2450680.

[2] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[3] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces for communication and rehabilitation," *Nature Rev. Neurol.*, vol. 12, no. 9, pp. 513–525, Sep. 2016, doi: 10.1038/nrneurol.2016.113.

[4] M. Hamedi, S.-H. Salleh, and A. M. Noor, "Electroencephalographic motor imagery brain connectivity analysis for BCI: A review," *Neural Comput.*, vol. 28, no. 6, pp. 999–1041, Jun. 2016, doi: 10.1162/NECO_a_00838.

[5] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001, doi: 10.1109/5.939829.

[6] J. Decety and D. H. Ingvar, "Brain structures participating in mental simulation of motor behavior: A neuropsychological interpretation," *Acta Psychologica*, vol. 73, no. 1, pp. 13–34, Feb. 1990.

[7] Y. Liu *et al.*, "Motor-imagery-based teleoperation of a dual-arm robot performing manipulation tasks," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 3, pp. 414–424, Sep. 2019, doi: 10.1109/TCDS.2018.2875052.

[8] R. Zhang *et al.*, "Control of a wheelchair in an indoor environment based on a brain–computer interface and automated navigation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 1, pp. 128–139, Jan. 2016, doi: 10.1109/TNSRE.2015.2439298.

[9] S. Bajaj, A. J. Butler, D. Drake, and M. Dhamala, "Brain effective connectivity during motor-imagery and execution following stroke and rehabilitation," *NeuroImage, Clin.*, vol. 8, pp. 572–582, 2015, doi: 10.1016/j.nicl.2015.06.006.

[10] L. M. Alonso-Valerdi, R. A. Salido-Ruiz, and R. A. Ramirez-Mendoza, "Motor imagery based brain–computer interfaces: An emerging technology to rehabilitate motor deficits," *Neuropsychologia*, vol. 79, pp. 354–363, Dec. 2015, doi: 10.1016/j.neuropsychologia.2015.09.012.

[11] S. D. Vries and T. Mulder, "Motor imagery and stroke rehabilitation: A critical discussion," *Acta Derm Venereol.*, vol. 39, no. 1, pp. 5–13, 2007, doi: 10.2340/16501977-0020.

[12] J. Luo, Z. Feng, J. Zhang, and N. Lu, "Dynamic frequency feature selection based approach for classification of motor imageries," *Comput. Biol. Med.*, vol. 75, pp. 45–53, Aug. 2016, doi: 10.1016/j.compbiomed.2016.03.004.

[13] R. Fu, Y. Tian, T. Bao, Z. Meng, and P. Shi, "Improvement motor imagery EEG classification based on regularized linear discriminant analysis," *J. Med. Syst.*, vol. 43, no. 6, p. 169, Jun. 2019, doi: 10.1007/s10916-019-1270-0.

[14] E. Dong, C. Li, L. Li, S. Du, A. N. Belkacem, and C. Chen, "Classification of multi-class motor imagery with a novel hierarchical SVM algorithm for brain–computer interfaces," *Med. Biol. Eng. Comput.*, vol. 55, no. 10, pp. 1809–1818, Oct. 2017, doi: 10.1007/s11517-017-1611-4.

[15] G. Pfurtscheller, C. Guger, and H. Ramoser, "EEG-based brain-computer interface using subject-specific spatial filters," in *Engineering Applications of Bio-Inspired Artificial Neural Networks*, vol. 1607, J. Mira and J. V. Sánchez-Andrés, Eds. Berlin, Germany: Springer, 1999, pp. 248–254.

[16] Q. Novi, C. Guan, T. H. Dat, and P. Xue, "Sub-band common spatial pattern (SBCSP) for brain-computer interface," in *Proc. 3rd Int. IEEE/EMBS Conf. Neural Eng.*, Kohala Coast, HI, USA, May 2007, pp. 204–207, doi: 10.1109/CNE.2007.369647.

[17] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Hong Kong, Jun. 2008, pp. 2390–2397, doi: 10.1109/IJCNN.2008.4634130.

[18] K. P. Thomas, C. Guan, C. T. Lau, A. P. Vinod, and K. K. Ang, "A new discriminative common spatial pattern method for motor imagery brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 11, pp. 2730–2733, Nov. 2009, doi: 10.1109/TBME.2009.2026181.

[19] W. Wu, X. Gao, and S. Gao, "One-versus-the-rest (OVR) algorithm: An extension of common spatial Patterns(CSP) algorithm to multi-class case," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, Shanghai, China, Jan. 2006, pp. 2387–2390, doi: 10.1109/IEMBS.2005.1616947.

[20] X. Xie, Z. L. Yu, H. Lu, Z. Gu, and Y. Li, "Motor imagery classification based on bilinear sub-manifold learning of symmetric positive-definite matrices," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 504–516, Jun. 2017, doi: 10.1109/TNSRE.2016.2587939.

[21] J. Luo, Z. Feng, and N. Lu, "Spatio-temporal discrepancy feature for classification of motor imageries," *Biomed. Signal Process. Control*, vol. 47, pp. 137–144, Jan. 2019, doi: 10.1016/j.bspc.2018.07.003.

[22] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Feb. 2018, doi: 10.1155/2018/7068349.

[23] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[24] Y. Li, X.-R. Zhang, B. Zhang, M.-Y. Lei, W.-G. Cui, and Y.-Z. Guo, "A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1170–1180, Jun. 2019, doi: 10.1109/TNSRE.2019.2915621.

[25] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3D convolutional neural network for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2164–2177, Oct. 2019, doi: 10.1109/TNSRE.2019.2938295.

[26] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, Nov. 2018, doi: 10.1109/TNNLS.2018.2789927.

[27] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision—ECCV*, vol. 9911, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 499–515.

[28] A. L. Guennec, S. Malinowski, and R. Tavenard, "Data augmentation for time series classification using convolutional neural networks," p. 9.

[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[33] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, p. 39, 2012, doi: 10.3389/fnins.2012.00039.

[34] R. Leeb, C. Brunner, G. R. Muller-Putz, and A. Schlogl, "BCI competition 2008—Graz data set B," p. 6.

[35] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain–computer communication: Motivation, aim, and impact of exploring a virtual apartment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 473–482, Dec. 2007, doi: 10.1109/TNSRE.2007.906956.

[36] Q. Zheng, F. Zhu, and P.-A. Heng, "Robust support matrix machine for single trial EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 551–562, Mar. 2018, doi: 10.1109/TNSRE.2018.2794534.

[37] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted Boltzmann machines," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 566–576, Jun. 2017, doi: 10.1109/TNSRE.2016.2601240.

[38] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, pp. 2579–2605, 2008.