PAPER

An efficient CNN-LSTM network with spectral normalization and label smoothing technologies for SSVEP frequency recognition

Journal of Neural Engineering

RECEIVED
22 May 2022

REVISED
12 August 2022

ACCEPTED FOR PUBLICATION
30 August 2022

PUBLISHED
12 September 2022

PAPER

An efficient CNN-LSTM network with spectral normalization and label smoothing technologies for SSVEP frequency recognition

Yudong Pan[1], **Jianbo Chen**[1], **Yangsong Zhang**[1,3,*] **and Yu Zhang**[2]

[1] Laboratory for Brain Science and Medical Artificial Intelligence,School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, People's Republic of China
[2] Department of Bioengineering, Lehigh University, Bethlehem, PA 18015, United States of America
[3] MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: zhangysacademy@gmail.com

Keywords: steady-state visual evoked potentials (SSVEPs), convolutional neural network (CNN), long short-term memory (LSTM), spectral normalization, label smoothing

Abstract

Objective. Steady-state visual evoked potentials (SSVEPs) based brain–computer interface (BCI) has received great interests owing to the high information transfer rate and available large number of targets. However, the performance of frequency recognition methods heavily depends on the amount of the calibration data for intra-subject classification. Some research adopted the deep learning (DL) algorithm to conduct the inter-subject classification, which could reduce the calculation procedure, but the performance still has large room to improve compared with the intra-subject classification. *Approach*. To address these issues, we proposed an efficient SSVEP DL NETwork (termed SSVEPNET) based on one-dimensional convolution and long short-term memory (LSTM) module. To enhance the performance of SSVEPNET, we adopted the spectral normalization and label smoothing technologies during implementing the network architecture. We evaluated the SSVEPNET and compared it with other methods for the intra- and inter-subject classification under different conditions, i.e. two datasets, two time-window lengths (1 s and 0.5 s), three sizes of training data. *Main results*. Under all the experimental settings, the proposed SSVEPNET achieved the highest average accuracy for the intra- and inter-subject classification on the two SSVEP datasets, when compared with other traditional and DL baseline methods. *Significance*. The extensive experimental results demonstrate that the proposed DL model holds promise to enhance frequency recognition performance in SSVEP-based BCIs. Besides, the mixed network structures with convolutional neural network and LSTM, and the spectral normalization and label smoothing could be useful optimization strategies to design efficient models for electroencephalography data.

1. Introduction

Brain–computer interface (BCI) is capable of translating user's intention by decoding electroencephalogram (EEG) to interact with the equipment [1]. The steady-state visual evoked potential (SSVEP) is one of the most popular EEG signals to build a BCI system which could yield high information transfer rate and provide enough operation commands [2]. Owing to its portability, low cost and excellent temporal resolution, EEG is the most widely imaging modality in BCI systems. EEG is prone to being contaminated by the various noises and artifacts. Thus, a robust algorithm that can quickly and accurately execute classification is vital for practical application that can deliver satisfactory performance. For different BCI paradigms, the different EEG analysis methods were proposed in past decades.

Although the SSVEP has explicit frequency spectrum characteristic that has the fundamental and harmonic frequencies as the flickering stimulus, the most popular methods extract the features and conduct the

classification based on time domain signals. In the literature, the methods based on multivariate statistical and spatial filtering algorithms are widely adopted and proposed [3]. A prevalent traditional recognition method is the canonical correlation analysis (CCA), which tries to find a pair of projection vectors to obtain maximum correlation coefficients between the test signals and reference signals [4]. Till now, there are many improved algorithms based on CCA, such as individual template canonical correlation analysis (IT-CCA) [5], multiway canonical correlation analysis [6], filter bank canonical correlation analysis [7], spatio-spectral canonical correlation analysis [8], etc. CCA and its variant methods can achieve satisfactory results on the SSVEP-based BCI system that has a small number of classification targets, but the performance drops rapidly when was implemented in the system has a large number of classification targets [9].

From 2018, the spatial filtering algorithms based on individual training dataset became dominant in the 40-class BCI systems, such as task-related component analysis (TRCA) [10], and correlated component analysis (CORRCA) [11], etc. Some variants based on TRCA and CORRCA were also proposed, such as latency aligning TRCA [12], two-stage CORRCA [13], etc. These methods need the calibration data to calculate the spatial filters for all the stimulus frequencies. In fact, the calculation procedure is time-consuming and laborious. Therefore, it is vital to improve the classification performance under a short calibration time. One solution could be developing better algorithms that need little calibration data. For instance, Wong *et al* developed a new scheme that applied to learning the data corresponding to not only the target stimulus but also other stimuli [14]. Jorajuría *et al* proposed a pipeline that extracts oscillatory sources and classifies them using the newly developed tensor-based discriminant analysis with shrinkage, which is robust even only a few calibration trials are available [15]. Georgiadis *et al* introduced a symbolic dynamics approach based on semi-supervised learning that uses templates and confidence intervals derived from a small training sample with labeled SSVEP responses [16]. Another way is to utilize the EEG data from existing subjects to facilitate the algorithm implementation for a new subject. For example, a cross-subject fusion method was proposed in [17] to efficiently shorten the calibration time of the target user using the data from other subjects. A cross-subject assistance framework was introduced in [18] to enhance the robustness of SSVEP recognition by maximizing inter- and intra-subject correlation.

Although the classical traditional methods have achieved considerable progress, they are highly dependent on handcrafted data feature extraction and calibration data. Benefited by the rapid development of deep learning (DL) in recent years, researchers began to develop the algorithms for EEG data analysis with DL models [19]. For the SSVEP data, several DL models have been proposed based on different angles. For instance, Waytowich *et al* used a compact convolutional neural network (Compact-CNN) named as EEGNet for inter-subject classification that yielded about 80% accuracy for a 12-class SSVEP dataset without the need for user-specific calibration [20]. To preserve the interpretability idea of solving correlation coefficient by conventional CCA method, Li *et al* devised a CCN-based non-linear mode, i.e. convolutional correlation analysis [21]. In order to make full use of the spectrum and phase information in SSVEP signal, Ravi *et al* utilize fast Fourier transformation (FFT) to convert the SSVEP signal from time domain to frequency domain and design a shallow CNN to process spectrum data [22]. Being that the effective information embedded in the harmonic frequency referring to target recognition, Ding *et al* have developed an algorithm to fuse features extracted by CNN in distinct frequency bands to improve performance [23]. To meet the real-time processing requirements with massive stimulus targets, Zhao *et al* proposed a multi-target fast classification method for augmented-reality-based SSVEP using CNN [24].

Owing to the large subject variabilities, the accuracy of inter-subject classification obtained by DL models is worse than those of intra-subject classification, although the sizes of training sample increased a lot. Meanwhile, DL models suffer from data insufficiency when were used for intra-subject classification. To address these issues, we proposed an efficient SSVEP DL NETwork (termed SSVEPNET) based on one-dimensional (1D) convolution and long short-term memory (LSTM) module. In the network, CNN is utilized to extract the spatio-temporal features of SSVEP, while LSTM encodes the spatio-temporal features according to the dependence between features, and finally these fine-grained features were input into three dense layers. To solve the possible over-fitting problem of SSVEPNET, we adopted the spectral normalization and label smoothing technologies to regularize the network during implementation. In order to verify the effectiveness of the proposed model, we evaluated our model on two datasets with intra-subject and inter-subject classification under different conditions, i.e. two datasets, two time-window lengths (1 s and 0.5 s), three sizes of training data. Experimental results show that our model outperforms other compared models in intra-subject and inter-subject classification. Furthermore, the ablation experiments and t-stochastic neighborhood embedding (t-SNE) visualization were conducted to further investigate the effectiveness, rationality and interpretability of our model design.

2. Related work

2.1. IT-CCA

CCA is a statistical analysis technique which finds the underlying correlation between two multivariate variables. Given two multidimensional variables $X \in \mathbb{R}^{n\times k}$ and $Y \in \mathbb{R}^{m\times k}$, CCA seeks to find a pair of projection vectors $W_x \in \mathbb{R}^{n\times 1}$ and $W_y \in \mathbb{R}^{m\times 1}$ to maximize the correlation between the two linear combinations $x = X^{\mathrm{T}}W_x$ and $y = Y^{\mathrm{T}}W_y$. These projection vectors and the correlation metric could be obtained by the following formula:

$$\begin{aligned}\rho(x,y) &= \underset{W_x,W_y}{\operatorname{argmax}} \frac{E(x^{\mathrm{T}}y)}{\sqrt{E(x^{\mathrm{T}}x)E(y^{\mathrm{T}}y)}} \\ &= \underset{W_x,W_y}{\operatorname{argmax}} \frac{E(W_X^{\mathrm{T}}XYW_Y)}{\sqrt{E(W_X^{\mathrm{T}}XX^{\mathrm{T}}W_X)E(W_Y^{\mathrm{T}}YY^{\mathrm{T}}W_Y)}}.\end{aligned} \tag{1}$$

In the CCA-based SSVEP recognition approaches, X is the multi-channel EEG data, and Y is the reference signal created with the sine-cosine function at some stimulation frequency. However, artificial reference signal Y is lack of the specific characteristics of subjects.

IT-CCA replaces the artificial reference signals by the individual template reference signal obtained by averaging multiple EEG trials of specific subject:

$$Y_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_j, j = 1,2,\ldots,K \tag{2}$$

where K represents the number of stimulation frequency, and N_i is number of trials of at frequency $f_i, i = 1,2,\ldots,K$. $X_j (j = 1,2,\ldots,N_i)$ is a single trial of EEG data. Y_i is the individual template. Then, for a test sample, its frequency is that of the reference signals with the maximum correlation:

$$f_{\text{target}} = \underset{f_i}{\operatorname{argmax}} \rho_i(x, Y_i), i = 1,2,\ldots,K. \tag{3}$$

2.2. TRCA

TRCA could extract task related components by maximizing the reproducibility of neuroimaging data in each task. It was first used for near-infrared spectroscopy data analysis [25]. In 2018, Nakanishi *et al* introduce TRCA for 40-class SSVEP data classification [10]. TRCA seeks to find a linear coefficient vector $w \in \mathbb{R}^{N_c\times 1}$ to maximize the inter-trial correlation of its projections $y(t) = w^{\mathrm{T}}x(t)$, which is called task-related components. Note that h_1th trial and task-related component are $x^{(h_1)}(t)$ and $y^{(h_1)}(t)$, h_2th trial and task-related component are $x^{(h_2)}(t)$ and $y^{(h_2)}(t)$,then the covariance between $y(t)$ of h_1th trial and h_2th trial can be expressed by the following formula:

$$\begin{aligned}C_{h_1,h_2} &= \operatorname{Cov}\left(y^{(h_1)}(t), y^{(h_2)}(t)\right) \\ &= \sum_{j_1,j_2=1}^{N_c} w_{j_1}w_{j_2}\operatorname{Cov}\left(x_{j_1}^{(h_1)}(t), x_{j_2}^{(h_2)}(t)\right).\end{aligned} \tag{4}$$

The combination of all trials can be expressed by the following formula:

$$\begin{aligned}\sum_{\substack{h_1\neq h_2\\ h_1,h_2=1}}^{N_t} C_{h_1,h_2} &= \sum_{\substack{h_1\neq h_2\\ h_1,h_2=1}}^{N_t}\sum_{j_1,j_2=1}^{N_c} w_{j_1}w_{j_2}\operatorname{Cov}\left(x_{j_1}^{h_1}(t), x_{j_2}^{h_2}(t)\right) \\ &= w^{\mathrm{T}}Sw\end{aligned} \tag{5}$$

where the symmetric matrix S can be expressed as:

$$S_{j_1,j_2} = \sum_{\substack{h_1\neq h_2\\ h_1,h_2=1}}^{N_t} \operatorname{Cov}\left(x_{j_1}^{(h_1)}(t), x_{j_2}^{(h_2)}(t)\right). \tag{6}$$

In order to get an effective solution, a normalization constraint is conducted on $y(t)$, i.e.

$$\begin{aligned}\operatorname{Var}(y(t)) &= \sum_{j_1,j_2=1}^{N_c} w_{j_1}w_{j_2}\operatorname{Cov}\left(x_{j_1}^{(h_1)}(t), x_{j_2}^{(h_2)}(t)\right) \\ &= w^{\mathrm{T}}Qw = 1.\end{aligned} \tag{7}$$

Finally, the constrained optimization problem is transformed into Rayleigh Ritz eigenvalue problem:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \frac{w^{\mathrm{T}}Sw}{w^{\mathrm{T}}Qw}. \tag{8}$$

TRCA obtains the corresponding spatial filter w_i, using w_i for test trial X and individual template signal $\overline{X}_i$ performs spatial filtering, and then calculates Pearson correlation coefficients:

$$r_i = \rho\left(X^{\mathrm{T}}w_i, \left(\overline{X}_i\right)^{\mathrm{T}}w_i\right). \tag{9}$$

Eventually, the stimulation frequency corresponding to the maximum correlation coefficient can be found:

$$f_{\text{target}} = \underset{f_i}{\operatorname{argmax}} r_i, i \in 1,2,\ldots,K. \tag{10}$$

In current study, to further improve the performance of TRCA, the extended TRCA with filter bank technology was used for classification. The selection of filter banks follows the design principles proposed by Chen *et al* [7]. Each sub-band component was extracted using zero phase Chebyshev type I infinite impulse response. The detailed sub-band ranges will be given in section 3.1 for the two used datasets.

2.3. DL-based methods

In the field of EEG research, regarding the network framework of DL, nearly 40% of the research uses CNN and nearly 14% of the research uses recurrent neural network [19]. This result shows that the network architecture based on DL has important potential and value for EEG research. For the SSVEP data analysis, there are some works using DL methods in recent years. Here, we briefly introduce three SSVEP models, which were adopted as the baseline methods.

- **EEGNet**: EEGNet is a compact model, and has become a baseline model for EEG data analysis [20, 26–28]. It mainly consists of a temporal convolution layer, a depthwise convolution, a separable convolution layer, and a fully connected layer. Because of the effectiveness, Waytowich *et al* utilize EEGNet for inter-subject classification on 12-class SSVEP dataset, and achieved about 80% accuracy [20].
- **FBtCNN**: Considering the important role of the harmonic component of the stimulation frequency in frequency recognition, Ding *et al* proposed a CNN architecture termed as FBtCNN that combines the time-domain signal as the input and uses the filter bank to fuse the characteristics of different frequency bands to improve the performance [23].
- **C-CNN**: The frequency domain of EEG data contains the spectrum and phase information related to the frequency recognition task. In order to make full use of the spectrum and phase information in SSVEP signal, Ravi *et al* utilize FFT to convert the SSVEP signal from time domain to frequency domain and design a shallow convolutional neural network named C-CNN to process spectrum data [22].

3. Methods and materials

3.1. Dataset

The proposed methods and other methods involved in this study were evaluated on two datasets; a public dataset—Dataset A [29] and a dataset used in previous study—Dataset B [30].

3.1.1. Dataset A

This dataset was an open access dataset provide in [29]. Ten healthy subjects were seated in a comfortable chair 60 cm from an liquid crystal display (LCD) monitor in a dim room. The stimuli was arranged in a 4×3 grid space as a simulation of telephone dialing interface. Twelve flickering stimulus (9.25 Hz:0.5 Hz:14:75 Hz) were displayed on the monitor. Ten healthy subjects were recruited to participate in the offline experiment. The EEG data were collected by the BioSemi ActiveTwo EEG system at a sampling rate of 2048 Hz. Eight Ag/AgCl electrodes PO7, PO3, POZ, PO4, PO8, O1, Oz and O2 were placed covering the occipital area. For each subject, they participated 15-block experiments. Each block contained 12 trials corresponding to all 12 stimuli generated in a random order. Each trial lasted for 5 s, which comprised of 1 s cuing period and 4 s targeted stimulus respectively. A total of 180 trials were collected. The more detailed information refers to [29].

All data was down-sampled to 256 Hz and band-pass filtered between 6 and 80 Hz via fourth-order band-pass forward-backward Butterworth bandpass filter. Considering the visual latency, we extract the data from the 135 ms after the stimulus onset. Two time windows were used, i.e. 0.5 s, 1.0 s, in the offline analysis. Four filter banks ranged from $[8 \times m, 80]$ Hz were leveraged in the filter bank technology based traditional methods, where $m \in \{1, 2, 3, 4\}$ represents mth sub-band.

3.1.2. Dataset B

In this dataset, ten healthy subjects with normal or corrected vision participated in the experiment. The experimental operations for this dataset were reviewed and approved by the ethics committee at the East China University of Science and Technology. Four red squares, respectively coded by 6 Hz, 8 Hz, 9 Hz, 10 Hz, were displayed in the middle of the four side lines of a square black screen.

EEG signals were acquired using the Nuamps amplifier with a sampling rate of 250 Hz. Thirty channels were placed on the standard positions according to the 10–20 international system. Each subject needed to complete 20 runs in the experiments. In each run, the subject was asked to gaze at each of four red squares for 4 s, which lead to a total of 80 trials in the whole experiments.

All data was band-pass filtered between 4 and 45 Hz via sixth-order band-pass forward-backward Butterworth bandpass filter. Since the occipital and parietal scalp areas have been demonstrated to contribute most to the recognition of SSVEP, we chose eight channels P7, P3, Pz, P4, P8, O1, Oz and O2 for analysis in this study. For this dataset, the data epochs were extracted after the stimulus onset. Two time windows were used, i.e. 0.5 s, 1.0 s, in the offline analysis. Resemble with Dataset A, four filter banks ranged from $[6 \times m, 45]$ Hz were employed in the filter bank technique based traditional methods.

3.2. The proposed CNN-LSTM model

The model structure of the proposed SSVEPNET is illustrated in figure 1. It mainly consists of five modules, an input layer, a spatial filtering module, a temporal filtering module, a bi-directional long short-term memory (Bi-LSTM) module, and a fully connected module. Our implementation is based on the PyTorch framework, so in the input layer, we require the data input format to be $N_c \times N_t$, where N_c denotes number of channels, N_t indicates the number of sample points. We used the spatial filtering module to learn $2 \times N_c$ spatial filters in the model. And we regularize each spatial filter by using a maximum

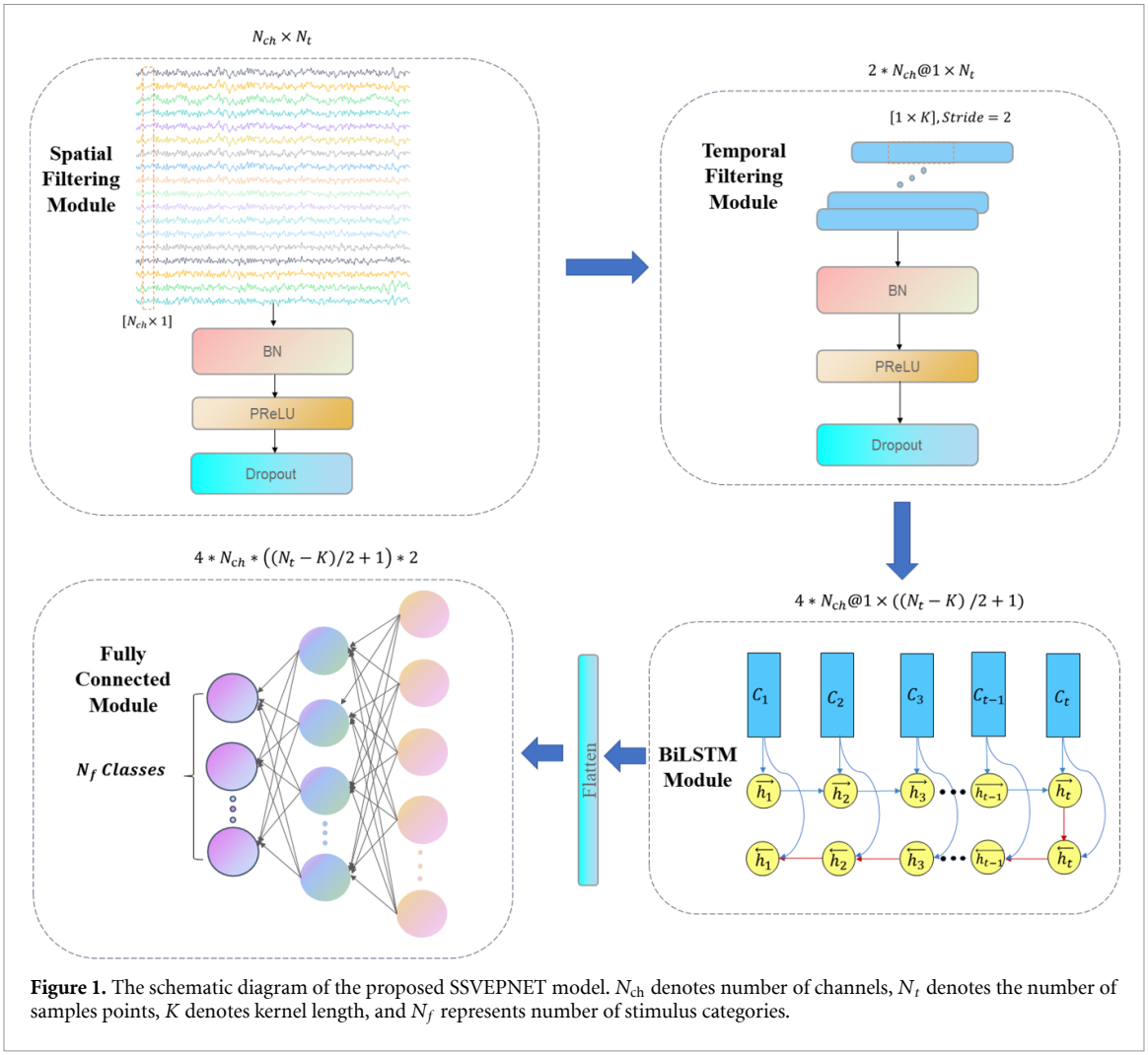

Figure 1. The schematic diagram of the proposed SSVEPNET model. N_{ch} denotes number of channels, N_t denotes the number of samples points, K denotes kernel length, and N_f represents number of stimulus categories.

norm constraint of 1 on its weights [26], i.e. $\|w\|^2 < 1$. Then, we extracted the temporal features of each group of signals after spatial filtering through temporal filter layer with kernel length equals K. In this study, we set K as 10. Due to the locality of the convolution kernel action, the network cannot learn global feature information through convolution kernels. Thus, a Bi-LSTM will help to learn the correlation and dependence between the spatial-temporal characteristics obtained by the first two-step operations. In order to learn the mapping relationship between these encoded fine-grained spatio-temporal features and classification targets, fully connected module composed of three dense layers were used. We expect that the dimension of spatio-temporal features is compressed and the relation between feature information and category information is more explicit after each dense layer.

It is worth noting that in order to ensure the nonlinearity of the mapping function learned by the convolution layers and the first two fully connected layers, a parametric rectified linear unit (PReLU) activation function was used after each layer. The softmax activation function was used in the last fully connected layer. Dropout layer with a probability of 0.5 between fully connected layers was used to avoid the potential over-fitting problem. Concretely, the detailed model parameters and training hyperparameters are displayed in table 1.

3.3. Attention-based label smooth

Label smooth was a technology to regularize the classifier layer via estimating the marginalized effect of label-dropout during training [33]. It is often used to reduce the overfitting problem of DL methods and further improve classification performance [34]. However, not much is known about why and when label smoothing should work [35]. In this study, we propose a label smoothing scheme based on visual attention mechanism. During the process of SSVEP data collection, the subject was guided to gaze at a flash stimulation through a clue. However, all the stimuli flickered simultaneously, the non-target stimuli are inevitably to distract the subject, which will also evoke the SSVEP components to disturb the frequency recognition. As shown in figure 2, the stimulus numbered 4 is the target stimulus, but the subject is likely influenced by the surrounding non-target stimuli, numbered 0, 1, 2, 3, 5, 6, 7, 8, respectively. In order to alleviate the adverse effects of this

Table 1. Hyperparameters of our model on two dataset.

	Dataset A		Dataset B	
Parameters	1 s	0.5 s	1 s	0.5 s
Spatial filtering kernel size	N_c	N_c	N_c	N_c
Temporal filtering kernel size	10	10	10	10
Temporal filtering kernel stride	2	2	2	2
Number of spatial filtering feature maps	$2 \times N_c$	$2 \times N_c$	$2 \times N_c$	$2 \times N_c$
Number of temporal filtering feature maps	$4 \times N_c$	$4 \times N_c$	$4 \times N_c$	$4 \times N_c$
Number of Bi-LSTM hidden size	$4 \times N_c$	$4 \times N_c$	$4 \times N_c$	$4 \times N_c$
Number of features of first layer (D1)	$8 \times N_c \times ((N_t - 10)/2 + 1)/10$	$8 \times N_c \times ((N_t - 10)/2 + 1)/10$	$8 \times N_c \times ((N_t - 10)/2 + 1)/10$	$8 \times N_c \times ((N_t - 10)/2 + 1)/10$
Number of features of second layer (D2)	D1/5	D1/5	D1/5	D1/5
Number of features of third layer (D3)	N_f	N_f	N_f	N_f
Learning rate	0.01	0.01	0.01	0.01
Max epochs	500	500	500	500
Mini batch-size	30	30	16	16
Drop probability	0.5	0.5	0.5	0.5
Weight decay	0.0003	0.0003	0.0003	0.0003
Activation function	PReLU [31]	PReLU	PReLU	PReLU
Optimizer (beta1, beta2)	Adam (0.9, 0.999)	Adam (0.9, 0.999)	Adam (0.9, 0.999)	Adam (0.9, 0.999)
Learning rate scheduler	Cosine annealing [32]	Cosine annealing	Cosine annealing	Cosine annealing

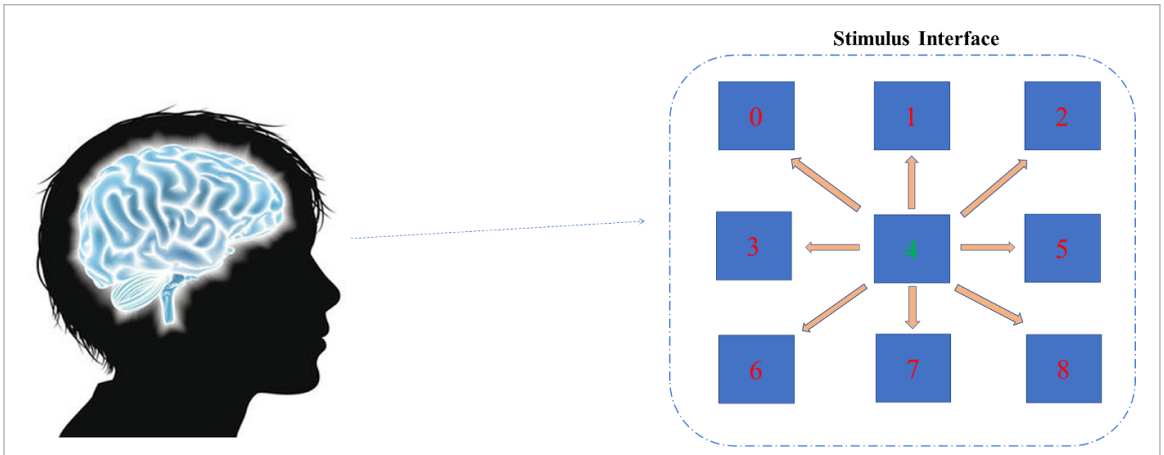

Figure 2. The schematic diagram of the proposed label smoothing scheme based on visual attention mechanism, in which the green number represents the correct stimulus target and red number represents the false stimulus target.

phenomenon, we extend the original single label to multiple labels.

More specifically, assuming that the K flickering stimuli were presented on an interface, and formed a stimulus matrix of size n rows by m columns. Then, for the flickering stimulus numbered as k, the coordinate (x, y) in the stimulus matrix is:

$$(x, y) = \left(\frac{k}{m}, k\%m\right) \tag{11}$$

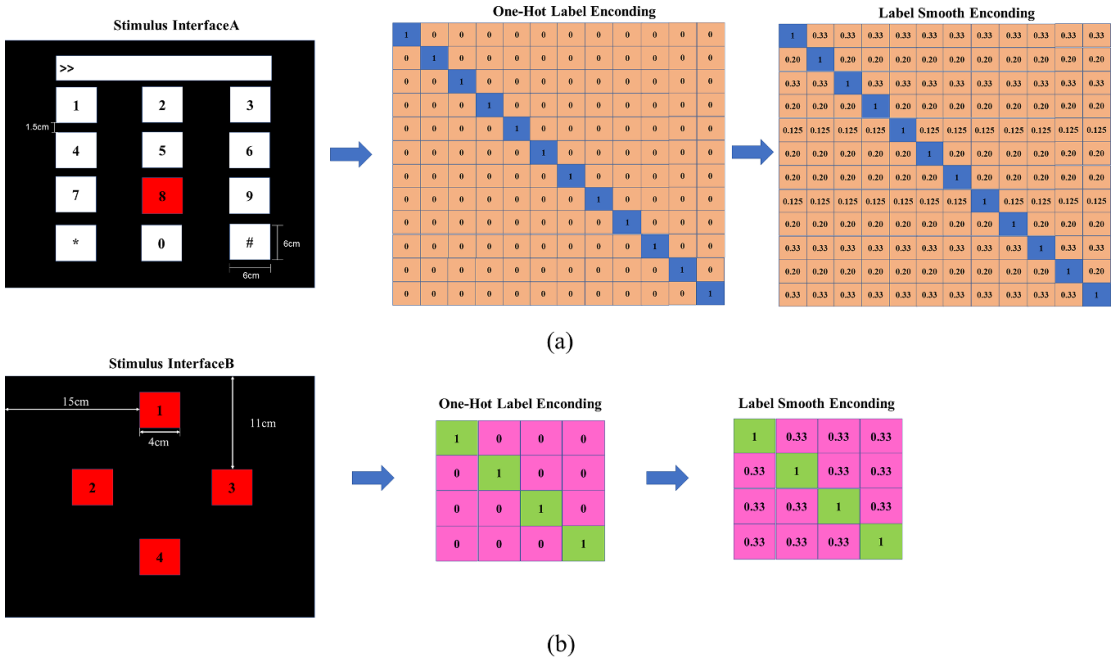

Figure 3. The diagram of one-hot coding scheme and label smoothing scheme for two datasets, each row in the matrices represents a label encoding vector. (a) Dataset A. (b) Dataset B.

where x indicates the row number, y indicates the column number. Suppose that the subjects' visual attention was evenly distracted by surrounding non-target stimuli of the gazed target stimulus, the attention scores for all the non-target stimuli was calculated as:

$$\beta_k = 1.0 \Big/ \sum_{\substack{i=-1 \\ i\neq j\neq 0}}^{1}\sum_{j=-1}^{1}((0 \leqslant (x+i) \leqslant n-1) \\ \&\&(0 \leqslant (y+j) \leqslant m-1)) \tag{12}$$

where β is attention score, $k \in [0, K-1]$ is the stimulation target number, && represents logical ADD operations. Then we obtain attention-based multi-label matrix:

$$\mathrm{ALS} = \begin{pmatrix} 1 & \beta_0 & \cdots & \beta_0 \\ \beta_1 & 1 & \cdots & \beta_1 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_k & \beta_k & \cdots & 1 \end{pmatrix} \tag{13}$$

where ALS represents the soft label matrix calculated based on the attention score, and each row in the matrix represents a label encoding vector. In order to avoid the excessive regularization of soft labels in the actual training process, which possibly makes the model difficult to converge, we utilize both hard labels and soft labels as supervision to train the model [34]

$$L = \alpha L_{\mathrm{hard}} + (1-\alpha)L_{\mathrm{soft}} \tag{14}$$

where L represents the total training loss, L_{hard} and L_{soft} represent the loss calculated using the cross-entropy under the one-hot coding label and ALS proposed in this study respectively. α is used to balance L_{hard} and L_{soft}. We only want this regularization technique to play an auxiliary role, thus α is set to 0.6 in this study. The schematic diagram of figure 3 simply illustrates the comparison between the label coding matrix generated by the label smoothing scheme proposed in this paper and the label coding matrix obtained based on the traditional one-hot coding method on Dataset A and Dataset B.

3.4. Spectral normalization

Spectral normalization has been widely used in adversarial generative networks and is often used in the discriminator to prevent gradient disappearance and gradient explosion, thus suppressing the mode collapse [36].

The central ideal of spectral normalization lies in limiting the updatable weight matrix in neural network to Lipschitz condition. Intuitively, the Lipschitz condition limits the severity of the change of the gradient of the function. In 1D space, it is easy to see that $y = \sin(x)$ is 1-Lipschitz, and its maximum slope is 1. To make weight matrix W satisfy K-Lipschitz continuous, the minimum value of K will be $\sigma(W) = \sqrt{\lambda_1}$, where λ_1 represents the maximum singular value for matrix $W^{\mathrm{T}}W$. Therefore, in order to limit weight matrix W satisfy 1-Lipschitz continuous and make the training process of the network more stable, we need to make the transformation for all elements in weight matrix W

$$\overline{W}_{\mathrm{SN}} := W/\sigma(W). \tag{15}$$

In light of the functional characteristics of spectral normalization, this regularization technology was

Table 2. The data portions of intra-subject experiments.

	Dataset A		Dataset B	
Data amount	Split ratio	Trial distribution	Split ratio	Trial distribution
Large	8:2	144:36	8:2	64:16
Middle	5:5	84:84	5:5	40:40
Small	2:8	36:144	2:8	16:64

introduced into the proposed SSVEPNET model. We adopted spectral normalization for each convolution layer and fully connected layer during implementing the network.

3.5. Experimental evaluation

To validate the effectiveness of our model, we adopted two classification scenarios, i.e. intra-subject and inter-subject classification. As illustrated in table 2, in the intra-subject classification scenario, the original data from each subject were divided into training set and testing set in the proportion of 8:2, 5:5 and 2:8 respectively. In terms of the size of training dataset, we marked these three sizes of datasets as large-scale, medium-scale and small-scale training dataset. Different sizes of datasets have impacts on the generalization performance of the model. The big dataset contains more heterogeneous samples and can be excavated more valuable knowledge, the models trained by large size of dataset that containing more different samples usually have better generalization and robustness [37]. However, collecting more data is time-consuming, and leads to visual fatigue. In order to eliminate the randomness of data division in the experiment, we adopted K-fold validation to alternately use each fold data as a testing dataset and the remaining folds as training datasets. The hyper-parameters of the network training under the intra-experimental paradigm can be found in table 1.

In the intra-subject experimental paradigm, the training data and test data are from the same subject, so the model trained in the intra-subject experimental paradigm is only applicable to a single subject. If we need to expand the training model to detect the data of subjects who do not participate in the training, we can adopt the inter-subject training strategy. Concretely, in the inter-subject classification, we use the leave-one subject out (LOSO) method to obtain the classification accuracy of each subject, and calculate the final average classification result of ten subjects as the final result.

For the DL models training, when the time-window length is 1 s, the hyper-parameters on two datasets were set as: mini-batch (B) = 64, max-epochs (E) = 500, learning rate (Lr) = 0.001, dropout probability (Dp) = 0.5. The optimizer we choose is Adam (beta1 = 0.9, beta2 = 0.999). When the time-window length is 0.5 s, the hyper-parameters of two datasets were set as: B = 30, E = 100, Lr = 0.01, Dp = 0.5, weight decay = 0.0001 (Dataset A)/0.0003 (Dataset B) and we utilize optimizer Adam (beta1 = 0.9, beta2 = 0.999) combined with cosine annealing to stable training.

In addition, it is worth mentioning that the results of the inter-subject experiment of each model are not very stable when the window length is 0.5 s, so we repeated the experiment five times based on the data division of the LOSO method, and the averaged accuracies were used as the final classification results.

3.6. Statistical analysis

In this study, the average classification results were presented with in the form of mean ± standard deviation. One-way repeated-measures analysis of variance (ANOVA) was used to test whether there were significant differences in the classification accuracy between these methods at each condition. Post hoc paired t-test was used to check whether there were significant differences between all pairs of methods at each condition. The alpha level was set at 0.05.

4. Results

To evaluate the effectiveness of the proposed model, we used two traditional methods (IT-CCA, TRCA), and three DL methods (EEGNet [20], C-CNN [22], FBtCNN [23]), as the compared baseline methods. For all these baseline methods, we implemented them according to the original studies.

4.1. The results of intra-subject classification experiments

In the intra-subject classification, SSVEPNET and all of other baseline traditional methods and DL methods are evaluated during the experiments. The classification results on two datasets at different conditions are elucidated in tables 3 and 4.

4.1.1. Dataset A

For the time window of 1 s, one-way repeated measures ANOVA revealed significant differences between all the classification methods under three different data divisions (small: $F(5,45) = 22.70$, $p < 0.001$, medium: $F(5,45) = 6.14$, $p < 0.001$, large: $F(5,45) = 8.36$, $p < 0.001$). In small-scale datasets, the classification methods can be ranked from lowest to highest as follows: FBtCNN, EEGNet, IT-CCA, C-CNN, TRCA, SSVEPNET. The post hoc paired t-tests showed that there were significant differences between the SSVEPNET and all of other methods (ITCCA: $p < 0.002$; TRCA: $p = 0.013$; EEGNet:

Table 3. Results of SSVEP classification on intra-subject experiments when time-window length is 1 s. The highest accuracy (in bold) indicates the optimal classification scheme for this scale of dataset.

	Dataset A			Dataset B		
Method	8:2	5:5	2:8	8:2	5:5	2:8
ITCCA	79.39 ± 19.08	77.56 ± 19.36	67.61 ± 24.88	75.12 ± 16.57	71.62 ± 17.08	66.06 ± 18.01
TRCA	97.28 ± 4.60	95.42 ± 8.24	83.28 ± 16.28	94.38 ± 6.62	93.25 ± 5.93	83.63 ± 13.21
EEGNet	91.27 ± 11.48	81.75 ± 20.94	55.28 ± 23.80	88.75 ± 10.88	76.88 ± 13.29	57.66 ± 12.00
FBtCNN	89.33 ± 15.08	81.00 ± 24.94	51.05 ± 20.52	92.25 ± 5.96	82.81 ± 11.71	64.34 ± 13.67
C-CNN	93.87 ± 8.82	86.33 ± 18.56	69.50 ± 21.86	95.50 ± 6.85	87.97 ± 10.05	76.91 ± 12.52
SSVEPNet	**98.20 ± 3.46**	**96.58 ± 6.05**	**88.62 ± 16.17**	**96.25 ± 6.07**	**95.16 ± 6.34**	**88.53 ± 10.65**

Table 4. Results of SSVEP classification on intra-subject experiments when time-window length is 0.5 s. The highest accuracy (in bold) indicates the optimal classification scheme for this scale of dataset.

	Dataset A			Dataset B		
Method	8:2	5:5	2:8	8:2	5:5	2:8
ITCCA	57.50 ± 14.84	52.74 ± 19.41	42.86 ± 19.11	56.12 ± 15.54	53.25 ± 13.81	47.28 ± 12.01
TRCA	90.56 ± 11.19	85.71 ± 14.96	66.44 ± 21.95	86.38 ± 9.23	83.75 ± 10.54	69.63 ± 14.83
EEGNet	79.93 ± 15.60	64.33 ± 18.77	35.63 ± 12.60	83.00 ± 8.72	71.72 ± 12.84	58.66 ± 12.44
FBtCNN	78.27 ± 19.26	68.42 ± 23.95	37.68 ± 16.34	85.00 ± 8.51	78.75 ± 10.32	56.16 ± 13.16
C-CNN	83.80 ± 15.07	78.50 ± 19.66	55.97 ± 20.30	89.88 ± 5.95	83.44 ± 5.64	70.41 ± 12.59
SSVEPNet	**92.20 ± 9.22**	**90.92 ± 10.81**	**79.23 ± 19.83**	**90.75 ± 5.22**	**90.47 ± 4.28**	**78.22 ± 11.38**

$p < 0.001$; C-CNN: $p < 0.001$, FBtCNN: $p < 0.001$). Similarly, in medium-scale datasets, SSVEPNET achieved higher performance than other methods. The post hoc paired t-tests showed that there were significant differences between the SSVEPNET and all of other methods (ITCCA: $p = 0.005$; TRCA: $p = 0.021$; EEGNet: $p = 0.021$; C-CNN: $p = 0.04$, FBtCNN: $p = 0.043$). However, on large-scale datasets, post hoc paired t-tests showed that there were marginally significant differences between SSVEPNET and C-CNN ($p = 0.054$), FBtCNN ($p = 0.059$), TRCA ($p > 0.05$), but there were significant differences between SSVEPNET and ITCCA ($p < 0.001$), EEGNet ($p = 0.041$).

For the time window of 0.5 s, the accuracies of all the models decrease obviously, but the SSVEPNET model yields better results. One-way repeated measures ANOVA manifested a significant difference between all the classification methods under three different data divisions (small-scale: $F(5,45) = 42.94$, $p < 0.001$, medium-scale: $F(5,45) = 35.36$, $p < 0.001$, large-scale: $F(5,45) = 36.77$, $p < 0.001$). Different from the condition when time window is 1 s, post hoc paired t-tests indicate that SSVEPNET significantly outperformed all of other methods under small-scale, medium-scale, large-scale datasets (ITCCA: $p < 0.001$, $p < 0.001$, $p < 0.001$; TRCA: $p < 0.002$, $p = 0.002$, $p = 0.008$; EEGNet: $p < 0.001$, $p < 0.001$, $p = 0.003$; C-CNN: $p < 0.001$, $p = 0.006$, $p = 0.007$; FBtCNN: $p < 0.001$, $p = 0.002$, $p = 0.006$), respectively.

4.1.2. Dataset B

In this dataset, the similar classification tendency was obtained as the Dataset A. Compared with other methods, the SSVEPNET had achieved the best classification results. For the time window of 1 s, one-way repeated measures ANOVA revealed a significant differences between all the classification methods under three different data divisions (small-scale: $F(5,45) = 39.68$, $p < 0.001$, medium-scale: $F(5,45) = 20.42$, $p < 0.001$, large-scale: $F(5,45) = 20.431$, $p < 0.001$). Post hoc paired t-tests indicated that SSVEPNET significantly outperformed all of other methods across small-scale dataset and medium-scale dataset (ITCCA: $p < 0.001$, $p < 0.001$; TRCA: $p < 0.005$, $p = 0.013$; EEGNet: $p < 0.001$, $p < 0.001$; C-CNN: $p < 0.001$, $p = 0.006$, FBtCNN: $p < 0.001$, $p < 0.001$), while there was no significant difference between SSVEPNET and the suboptimal method C-CNN ($p = 0.443$) and TRCA ($p = 0.052$) on large scale dataset.

Furthermore, when the time window is 0.5 s, the results are similar to those at 1 s time window. One-way repeated measures ANOVA revealed a significant difference between all the classification methods under three different data divisions (small-scale: $F(5,45) = 40.32$, $p < 0.001$, medium-scale: $F(5,45) = 34.742$, $p < 0.001$, large-scale: $F(5,45) = 49.134$, $p < 0.001$). Post hoc paired t-tests indicate that SSVEPNET significantly outperformed all of other methods under small-scale dataset and medium-scale dataset (ITCCA: $p < 0.001$, $p < 0.001$; TRCA: $p = 0.002$, $p = 0.003$; EEGNet: $p < 0.001$, $p < 0.001$; C-CNN: $p = 0.004$, $p < 0.001$, FBtCNN: $p < 0.001$, $p = 0.001$). And there was no significant difference between SSVEPNET and the suboptimal method C-CNN ($p = 0.435$), but there were significant differences between SSVEPNET and rest of other methods under large-scale dataset (ITCCA:

Table 5. The inter-subject classification results when time-window length is 1 s. The highest accuracy (in bold) indicates the optimal classification scheme for this subject.

	Dataset A				Dataset B			
Subject	EEGNet	FBtCNN	C-CNN	SSVEPNet	EEGNet	FBtCNN	C-CNN	SSVEPNet
S01	52.34	46.09	57.81	**58.59**	93.75	**100.00**	96.88	98.44
S02	44.53	40.62	36.72	**45.31**	**60.94**	51.56	54.69	51.56
S03	66.41	44.53	60.94	**71.88**	**68.75**	46.88	67.19	59.38
S04	**96.88**	92.19	95.31	96.09	68.75	87.50	73.44	**90.62**
S05	89.06	89.94	60.62	**92.19**	87.50	95.31	98.44	**100.00**
S06	**95.31**	88.28	95.31	94.53	95.31	98.44	93.75	**100.00**
S07	93.75	85.16	89.84	**95.31**	75.00	73.44	76.56	**87.50**
S08	97.66	96.09	96.88	**98.44**	82.81	76.56	81.25	**85.94**
S09	87.50	92.19	97.66	**97.66**	**81.25**	78.12	81.25	78.12
S10	84.38	77.56	80.47	**94.53**	**93.75**	87.50	95.31	90.62
Mean	80.78 ± 18.38	75.55 ± 21.23	80.70 ± 20.09	**84.45 ± 18.01**	80.78 ± 11.46	79.53 ± 17.46	81.88 ± 13.70	**84.22 ± 15.86**

Table 6. The inter-subject classification results when time-window length is 0.5 s. The highest accuracy (in bold) indicates the optimal classification scheme for this subject

	Dataset A				Dataset B			
Subject	EEGNet	FBtCNN	C-CNN	SSVEPNet	EEGNet	FBtCNN	C-CNN	SSVEPNet
S01	**34.89**	27.89	30.78	31.89	82.67	84.00	80.67	**84.33**
S02	**22.56**	17.00	17.22	18.78	39.00	30.67	**40.67**	39.67
S03	45.78	28.33	40.89	**47.11**	53.67	33.00	**57.33**	50.00
S04	**79.44**	68.89	71.33	74.22	71.00	66.00	71.00	**77.33**
S05	72.11	70.78	71.78	**82.11**	81.00	**89.00**	85.33	85.67
S06	**80.33**	74.33	75.11	77.33	73.67	70.67	79.33	**80.33**
S07	81.22	72.67	74.00	**85.00**	**66.00**	59.00	64.67	65.33
S08	79.33	73.89	78.78	**84.44**	56.33	59.67	56.00	**61.67**
S09	64.56	57.44	65.89	**69.22**	65.33	71.33	68.00	**76.67**
S10	57.22	52.78	50.44	**63.00**	**89.00**	58.00	80.00	76.00
Mean	61.74 ± 19.99	54.40 ± 20.93	57.62 ± 20.40	**63.31 ± 22.03**	67.77 ± 14.35	62.13 ± 18.01	68.30 ± 13.26	**69.70 ± 14.50**

$p < 0.001$; TRCA: $p = 0.027$; EEGNet: $p < 0.002$; FBtCNN: $p = 0.01$).

4.2. The results of inter-subject classification experiments

In the inter-subject classification, we cannot use the data of the target subjects, and thus difficult to calculate the effective template signals [38]. Therefore, in this part, we only compare our model with the three baseline DL models, i.e. EEGNet, FBtCNN and C-CNN. The classification results of ten subjects on two datasets and two different time windows (1.0 s, 0.5 s) are summarized in tables 5 and 6.

4.2.1. Dataset A

As shown in table 5, when the time window is 1 s, the classification results achieved by SSVEPNET was 84.45 ± 18.01%, which outperformed the other three DL methods (EEGNet: 80.78 ± 18.38%, FBtCNN: 75.55 ± 21.23%, C-CNN: 80.70 ± 20.09%). One-way repeated measures ANOVA revealed a significant difference between the four classification methods ($F(3, 27) = 7.882$, $p < 0.001$). When the time length is 0.5 s, as shown in table 6, the average results of our model (63.31 ± 22.03) for all subjects still outperformed other methods (EEGNet: 61.74 ± 19.99%, FBtCNN: 54.40 ± 20.93%, C-CNN: 57.62 ± 20.40%). One-way repeated measures ANOVA revealed significant differences between all the classification methods ($F(3, 27) = 16.35$, $p < 0.002$).

4.2.2. Dataset B

As the results in Dataset A, SSVEPNET achieves best classification performance at two different time windows (84.22 ± 15.86%, 69.70 ± 14.50%), and approximately 2.4% and 1.4% higher than that of the suboptimal algorithm C-CNN (81.88 ± 13.70%, 68.30 ± 13.26%) at 1 s and 0.5 s, respectively. However, one-way repeated measures ANOVA revealed that there were no significant differences between all the classification methods at 1 s ($F(3, 27) = 2.10$, $p < 0.2$), but there were significant differences at 0.5 s ($F(3, 27) = 7.04$, $p < 0.02$).

4.3. Ablation experiments

The DL model with large parameters is extremely prone to over-fitting when the amount of training data is small [37]. There are three types of strategies to suppress the over-fitting phenomenon of the model: (a) the strategy worked on model parameters or architecture, such as dropout, batch normalization and weight decay; (b) the strategy worked on model input, such as data augmentation methods or data corruption; (c) the strategy worked on model output, like label smoothing [39]. In this study, we utilized two strategies, i.e. model

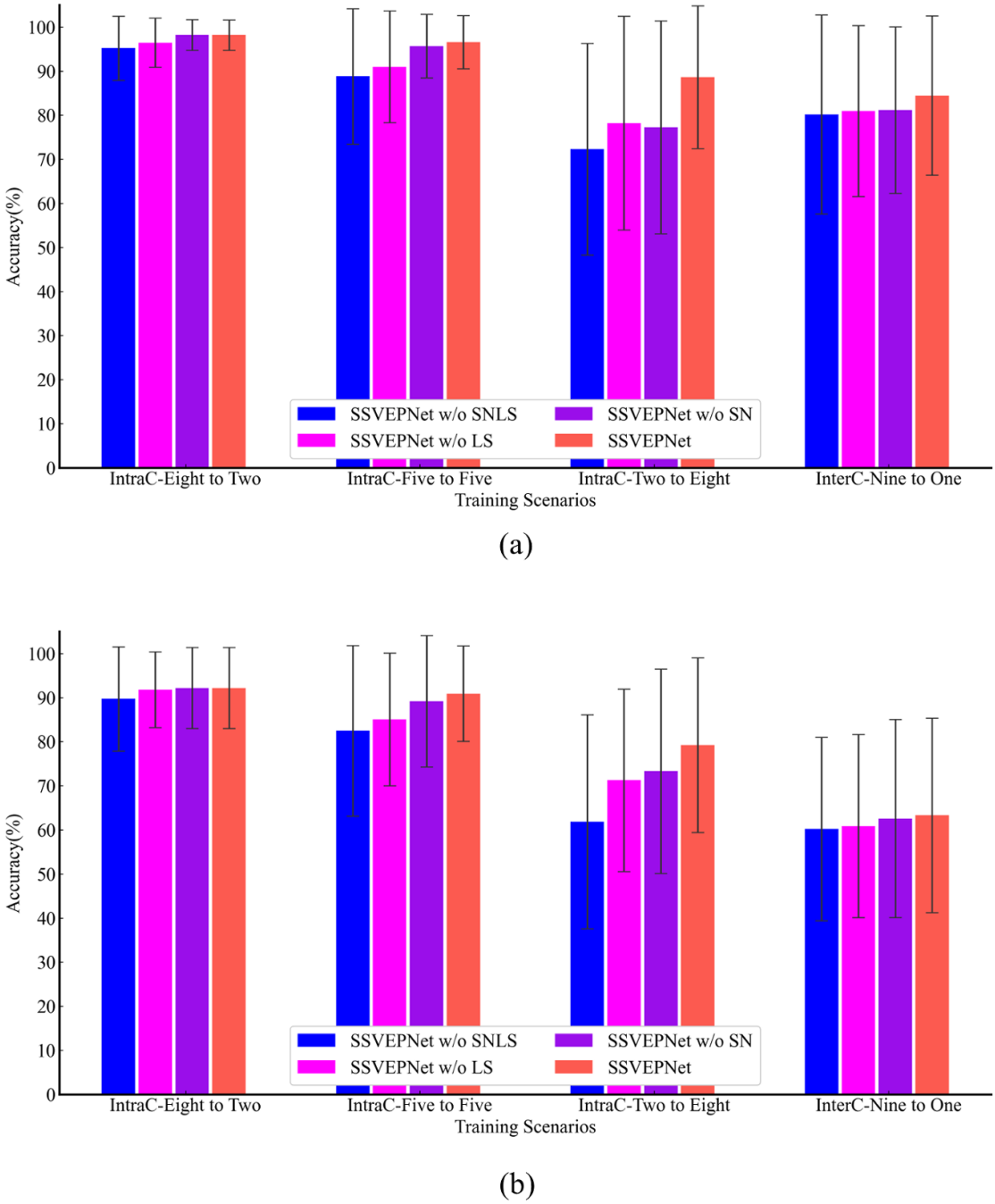

Figure 4. The results of ablation experiments on Dataset A. (a) The time window is 1 s, (b) the time window is 0.5 s. The horizontal axis represents the different ways of dividing the dataset, in which IntraC and InterC represent the intra-subject and inter-subject classification experiments.

structure (spectral normalization) and model output (attention-based label smoothing), to regularize the network. In order to verify the indispensability of the two strategies, we carried out ablation experiments on Dataset A. Specifically, we compared the classification results of the SSVEPNET model with or without the spectral normalization and label smoothing in intra-subject and cross-subject classification. We denoted the model that removing spectral normalization in SSVEPNET, the model that removing label smoothing in SSVEPNET, and the model that removing the spectral normalization and label smoothing as SSVEPNET-w/o-SN, SSVEPNET-w/o-LS, and SSVEPNET-w/o-SNLS, respectively. Figures 4(a) and (b) exhibit the ablation experimental results when the time window is 1 s and 0.5 s. It can be inferred that with the decreasing of the amount of training data, the role of the two regularization strategies become more beneficial, especially on the small-scale dataset in the case of intra-subject classification. The classification accuracy of the four models are 72.28 ± 24.01%, 78.20 ± 24.26%, 77.28 ± 24.14%, 88.62 ± 17.17 (1 s) and 61.85 ± 22.48%, 71.28 ± 20.74%, 73.33 ± 23.21%, 79.23 ± 19.83%

(0.5 s). One-way repeated measures ANOVA revealed a significant difference between all the classification methods (1 s: $F(3,27) = 9.08$, $p < 0.01$; 0.5 s: $F(3,27) = 40.92$, $p < 0.001$). In addition, it can be found that the two regularization techniques could achieve better results than that when only one of them was used. These results indicate that we can combine these regularization strategies to make the model achieve better generalization performance during developing DL models for SSVEP classification.

5. Discussion

It is still challenging to develop high-performance algorithms in BCI systems because of the nonstationary and nonlinear characteristics of EEG data. For SSVEP data, lots of traditional and DL methods were proposed in past years, and achieved great progress, such as CCA, TRCA, CORRCA, EEGNET, etc. For the methods as TRCA and CORRCA, they need calibration data, which is time-consuming and laborious to collect. Therefore, reducing the collection time, namely the data acquiring time, for a subject becomes a hot topic in the BCI field in recent years. One possible way is to leverage the data from the existing subjects to train a model for a specific subject. DL methods have received increasing attentions to design schemes. In current study, the DL methods yield better performance than the methods like TRCA, especially in the inter-subject classification owing to the large variants between subjects. In this study, we proposed a new mixed CNN-LSTM model to improve the classification results on 12-class and 4-class paradigms for SSVEP-based BCI. Our model achieved best results compared with other state-of-the-art (SOTA) methods under different evaluation experiments, which holds the potential to reduce the calibration time. To further unearth the reasons behind the high performance of SSVEPNET, here, we discuss the potential value and innovative perspective of our model for the design of SSVEP BCI system based on DL, and analyze its current limitations and explore the future research direction.

5.1. Potential benefits of ALS for DL methods with time-domain input

As one of the most important components of SSVEPNET, ALS plays an indispensable role in enhancing the performance. ALS is an improved regularization technology based on the soft label proposed by Szegedy *et al* [33], and the characteristics of targets' arrangement. Its main function is to suppress the influence of noise labels around the target on classification results. To further explore the functional utility of ALS, we conducted experiments to check the classification performance of the four models with and without the ALS. Based on the small-scale dataset of Dataset A and Dataset B, the classification results with 1 s time window are shown in table 7. For the SSVEPNET, EEGNet and FBtCNN, the classification accuracies were greatly improved with ALS (Dataset A: $p = 0.011$, $p < 0.001$, $p = 0.024$; Dataset B: $p < 0.001$, $p < 0.001$, $p = 0.121$). However, for the C-CNN model with complex spectrum features as input, the classification accuracies decreased with ALS (Dataset A: $p = 0.003$; Dataset B: $p < 0.001$). We speculate that when the SSVEP signal is evoked by multiple stimuli, the time-domain input signal can reflect the trajectories of multiple stimulus components over time, while the frequency-domain input cannot reflect this characteristic accounting for its statics. Hence, the network of time-domain input could capture this dynamic mode feature with ALS, to further improve performance. As for frequency-domain input network, adding ALS makes it more difficult to converge. In the future, we need to further verify the ALS on more DL methods and other BCI datasets.

5.2. Impact of number of parameters on the performance of SSVEP-BCI based on DL methods

In DL, the amount of network parameters has an important impact on the performance of the network. Generally speaking, the network with large parameters has stronger learning ability and is capable of fitting more complex nonlinear mapping functions, but it also needs more data to fit the network. Otherwise, it is easy to fall into over-fitting phenomenon and damage the generalization performance of the model. Nevertheless, as we all know, collecting a large amount of effective data is time-consuming and laborious in the BCI field [37]. In order to strive to retain the learning ability of the network with high number of parameters and avoid over-fitting simultaneously, there are some feasible schemes such as model compression [40] or designing a light network with more effective structure [41]. However, compressed model has more advantages in application deployment, but some compression techniques like structural matrix or transferred convolutional filters impose prior human knowledge, which could significantly affect the performance and stability to the model [40]. For the lightweight models, their representation ability is sometimes limited, especially for the complex data such as non-stationary and nonlinear EEG data. In this study, we combined the regularization technology and design an effective DL model with large-scale parameter to find the mapping relationship between the SSVEP data and the corresponding target labels. As shown in table 8, our network parameters have the largest amount of network parameters across two datasets and two time windows. Especially when the time-window length is 1 s on Dataset A, our model parameters reach 6443.87 K, which is about 140 times higher than the second highest model EEGNet and 673 times higher than the lowest model FBtCNN. Surprisingly, our model has

Table 7. Results of ablation experiments for attention-based label smooth technology. The highest accuracy (in bold) indicates the optimal scheme for this network.

		SSVEPNet	EEGNet	C-CNN	FBtCNN
Dataset A	w/o ALS	78.20 ± 24.26	55.28 ± 23.80	**69.50 ± 21.86**	51.05 ± 21.52
	w/ ALS	**88.62 ± 16.17**	**67.98 ± 21.74**	62.77 ± 22.49	**55.47 ± 20.63**
Dataset B	w/o ALS	74.78 ± 13.33	57.66 ± 12.00	**76.91 ± 12.52**	64.34 ± 13.67
	w/ ALS	**88.53 ± 10.05**	**67.50 ± 13.02**	64.84 ± 12.01	**68.69 ± 13.67**

Table 8. The amount of network parameters on two dataset. The network with the highest number of parameters under this time window is marked in bold.

	Dataset A		Dataset B	
Method	1 s	0.5 s	1 s	0.5 s
EEGNet	45.90 K	28.24 K	39.69 K	24.39 K
FBtCNN	9.57 K	2.91 K	9.24 K	2.84 K
C-CNN	43.31 K	43.31 K	16.29 K	16.29 K
SSVEPNet	**6443.87 K**	**1527.60 K**	**6137.07 K**	**1427.82 K**

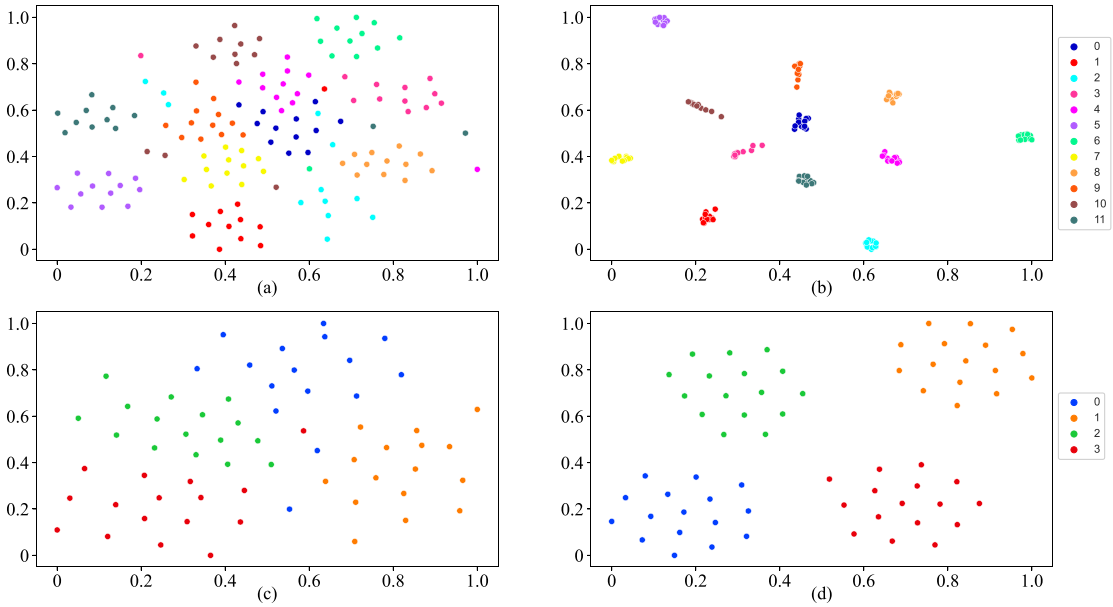

Figure 5. Feature distribution visualization of a representative subject (subject 8 on Dataset A, subject 5 on Dataset B) using tSNE for C-CNN and SSVEPNet at 1 s time window on the small-scale dataset in the intra-subject experiments. (a) C-CNN on Dataset A. (b) SSVEPNet on Dataset A. (c) C-CNN on Dataset B. (d) SSVEPNet on Dataset B.

better classification performance than other models in the case of less data and short time-window length as shown in tables 3–6. This may imply that we should rethink the idea of designing model architecture for SSVEP signal and other biomedical signal classification, to further break through the performance bottleneck of the network.

In the previous section, we discussed the impact of different parameters on network performance, and analyzed the difference of parameters between our model and other DL models. However, we did not explore the essential reason why our model is superior to other methods. Therefore, in this section, t-SNE technology will help us visualize the high-level features extracted from the DL model in a two-dimensional plane, thereby indirectly reflecting the distribution differences of these high-dimensional features in the multi-dimensional space [42]. Figure 5 illustrates the differences of feature distribution extracted by C-CNN and SSVEPNet at 1 s time window on the small-scale dataset in the intra-subject experiments. The data from two representative subjects 8 and 5 was used in Dataset A and Dataset B as comparison. Among them, figures 5(a) and (b) are the visualization results of feature distribution on Dataset A, while the visualization results of feature distribution on Dataset B are displayed in figures 5(c) and (d). It can be observed that compared with the C-CNN method, the between-class distances in the feature distribution extracted by SSVEPNET are larger and the within-class distances are smaller, which results in better classification performance.

In the EEGNet, FBtCNN and C-CNN, only one fully connected layer was used to learn the mapping

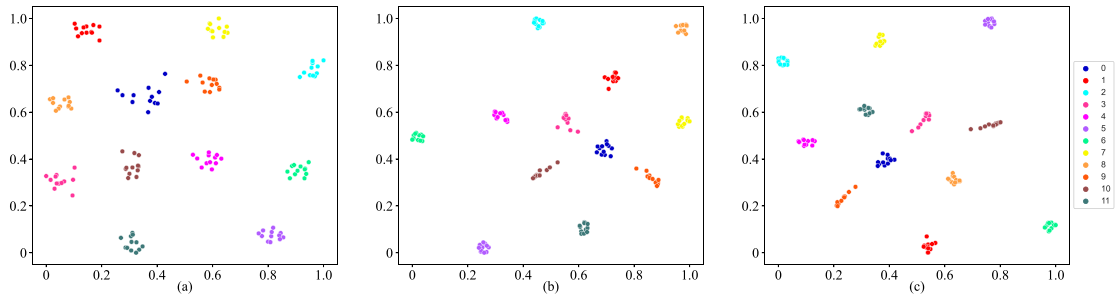

Figure 6. Feature distribution visualization in three fully connected layers of a representative subject (subject 8) on Dataset A using tSNE for SSVEPNet in the intra-subject experiments. (a) Input features of the first dense layer. (b) Input features of the second dense layer. (c) Input features of the third dense layer.

relationship between the features extracted by the convolution layers and the category information. However, this design idea is not applicable to the SSVEPNET network, it will be difficult to fit the corresponding mapping relationship between the high-dimensional features (7936) extracted by our model and the low-level category information. In addition, it will be unable to dig out more characteristics of data features to achieve better classification performance. In order to better verify the hypothesis, we utilize t-SNE technology to visualize the input features from the three fully connected layers in the fourth module in figure 1, as shown in figure 6. It can be observed that the features become more and more clustered as the features are processed by deeper fully connected layers. This implies that the fully connected layers can further mine more discriminative features based on the features extracted by convolution and LSTM layers.

5.3. Computation efficiency and superiority of SSVEPNET on small sample size data

Owing to the large amount of parameters, we need to check whether the SSVEPNET can be effectively used for online recognition. We calculated the computational time that the SSVEPNET conducted the classification on each trial at 1 s and 0.5 s, respectively. The computational time was evaluated with Pytorch 1.9.0 on a desktop computer with a NVIDIA GeForce GTX 1650 GPU (4 GB Memory). The averaged time of SSVEPNET model was 0.083 ms and 0.125 ms at 0.5 s and 1 s time windows, respectively. Accordingly, the SSVEPNET has low time computational complexity, and it may be a good candidate for the BCI community to use in their BCI applications.

In recent years, researchers have proposed some transfer learning strategies to improve inter-subject classification accuracy, which could reduce the user-dependent calibration data. For instance, Yuan *et al* calculated the SSVEP templates for a new subject by using the data from the existing subjects to enhance the detection of SSVEPs [43]. Chiang *et al* boosted the template-based SSVEP decoding algorithms through incorporating a least-squares transformation-based transfer learning method to leverage calibration data across multiple domains [44]. Wong *et al* introduced a subject transfer CCA to transfer the knowledge within subject and between subjects, which achieved high performance even the subject's calibration trials are fewer than the number of visual stimuli [45]. Discovering that the spatial filter and impulse response share commonality across neighboring stimulus frequencies, Wong *et al* developed a transfer learning CCA to further reduce calibration time [46]. In addition, a cross-subject spatial filter transfer method was proposed in [47], which transfers the existing user model with good SSVEP response to the new user test data without collecting calibration data. Although the above methods can achieve good performance, most of them rely on hand-crafted feature extraction, and some specific assumptions. In contrast, the proposed SSVEPNET model could provide an end-to-end classification scheme with little preprocessing operation. Besides, we could find that the SSVEPNET model was used for the subject-dependent and subject-independent scenarios, and achieved better results than the compared methods.

5.4. Limitation and future work

Some limitations for current study should be mentioned. We just evaluated the proposed models and achieved better results on 12-class and 4-class SSVEP offline datasets. In the future, we need to implement SSVEPNET in the online system for real decision processes. And, we also need to further verify the applicability of SSVEPNET on other datasets with large number of targets, such as Benchmark dataset [9]. In the inter-subject classification, when the time window is 0.5 s, the classification result became worse, we need to further explore more strategies to enhance the model performance with short time windows, such as data generation with the generative adversarial networks [48], domain adaptation technology [49]. Furthermore, despite of the low computation time and high classification performance, there is a significant difference between the laboratory environment and the real-world. For the applications out of the

laboratory, the methods having a great performance at the 1 s time window could ensure more robust performance. For instance, a system that leads to a decision every second could be great for communication for people with severe disabilities. Moreover, due to the complexity and non-stationarity of EEG data, the data collected by BCI users in different time periods could be significantly different as well [44]. When the data from different periods was input the BCI system, a regular break to retrain the network could be useful to yield better results. And, the adaptive strategy can be used to drop old data and update by current session data [50]. The transfer learning techniques should be adopted to take full advantage the old data to transfer to current session data [44, 51]. Besides, when the data from a new subject are available, that are not originally present in the database, incremental learning could be an effective technology to extend previously acquired knowledge [52, 53].

6. Conclusion

When the data amount is insufficient and the time-window length is short, designing a model with high classification accuracy is an urgent demand of SSVEP-based BCI system. In this study, we proposed an efficient mixed CNN-LSTM network with spectral normalization and label smoothing technologies for short-time SSVEP classification. To verify the effectiveness of our model, we comprehensively compared the proposed model with other traditional and DL methods both in intra-subject and inter-subject classification scenarios, and two different time-window lengths (1 s and 0.5 s). The results show that the proposed model is superior to other methods, especially on the small-scale dataset in intra-subject classification scenario. Further ablation experiments examined the role of spectral normalization and attention-based label smoothing in the proposed model. The extensive experimental results demonstrate that the SSVEPNET could be a promising candidate to achieve satisfactory performance for SSVEP frequency recognition, and spectral normalization and attention-based label smoothing could be a potential strategy.

Data availability statement

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

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62076209. We thank Ms Huan Cai for valuable comments on an early version of this manuscript.

ORCID iD

Yudong Pan https://orcid.org/0000-0002-3209-4721

References

[1] Wolpaw J R *et al* 2000 Brain-computer interface technology: a review of the first international meeting *IEEE Trans. Rehabil. Eng.* **8** 164–73
[2] Chen Y, Yang C, Ye X, Chen X, Wang Y and Gao X 2021 Implementing a calibration-free SSVEP-based BCI system with 160 targets *J. Neural Eng.* **18** 046094
[3] Zhang Y, Xu P, Cheng K and Yao D 2014 Multivariate synchronization index for frequency recognition of SSVEP-based brain-computer interface *J. Neurosci. Methods* **221** 32–40
[4] Lin Z, Zhang C, Wu W and Gao X 2006 Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs *IEEE Trans. Biomed. Eng.* **53** 2610–4
[5] Bin G, Gao X, Wang Y, Li Y, Hong B and Gao S 2011 A high-speed BCI based on code modulation VEP *J. Neural Eng.* **8** 025015
[6] Zhang Y, Zhou G, Zhao Q, Onishi A, Jin J, Wang X and Cichocki A 2011 Multiway canonical correlation analysis for frequency components recognition in SSVEP-based BCIs *Int. Conf. on Neural Information Processing* pp 287–95
[7] Chen X, Wang Y, Gao S, Jung T-P and Gao X 2015 Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface *J. Neural Eng.* **12** 046008
[8] Cherloo M N, Amiri H K and Daliri M R 2022 Spatio-spectral CCA (SS-CCA): a novel approach for frequency recognition in SSVEP-based BCI *J. Neurosci. Methods* **371** 109499
[9] Wang Y, Chen X, Gao X and Gao S 2016 A benchmark dataset for SSVEP-based brain–computer interfaces *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 1746–52
[10] Nakanishi M, Wang Y, Chen X, Wang Y-T, Gao X and Jung T-P 2017 Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis *IEEE Trans. Biomed. Eng.* **65** 104–12
[11] Zhang Y, Guo D, Li F, Yin E, Zhang Y, Li P, Zhao Q, Tanaka T, Yao D and Xu P 2018 Correlated component analysis for enhancing the performance of SSVEP-based brain-computer interface *IEEE Trans. Neural Syst. Rehabil. Eng.* **26** 948–56
[12] Huang J, Yang P, Xiong B, Wan B, Su K and Zhang Z-Q 2022 Latency aligning task-related component analysis using wave propagation for enhancing SSVEP-based BCIs *IEEE Trans. Neural Syst. Rehabil. Eng.* **30** 851–9
[13] Zhang Y, Yin E, Li F, Zhang Y, Tanaka T, Zhao Q, Cui Y, Xu P, Yao D and Guo D 2018 Two-stage frequency recognition method based on correlated component analysis for SSVEP-based BCI *IEEE Trans. Neural Syst. Rehabil. Eng.* **26** 1314–23
[14] Wong C M, Wan F, Wang B, Wang Z, Nan W, Lao K F, Mak P U, Vai M I and Rosa A 2020 Learning across multi-stimulus enhances target recognition methods in SSVEP-based BCIs *J. Neural Eng.* **17** 016026
[15] Jorajuría T, Idaji M J, İşcan Z, Gómez M, Nikulin V V and Vidaurre C 2022 Oscillatory source tensor discriminant analysis (OSTDA): a regularized tensor pipeline for SSVEP-based BCI systems *Neurocomputing* **492** 664–75
[16] Georgiadis K, Laskaris N, Nikolopoulos S and Kompatsiaris I 2018 Discriminative codewaves: a symbolic dynamics approach to SSVEP recognition for asynchronous BCI *J. Neural Eng.* **15** 026008
[17] Sun Y, Ding W, Liu X, Zheng D, Chen X, Hui Q, Na R, Wang S and Fan S 2022 Cross-subject fusion based on time-weighting canonical correlation analysis in SSVEP-BCIs *Measurement* **199** 111524

[18] Wang H, Sun Y, Wang F, Cao L, Zhou W, Wang Z and Chen S 2021 Cross-subject assistance: inter-and intra-subject maximal correlation for enhancing the performance of SSVEP-based BCIs *IEEE Trans. Neural Syst. Rehabil. Eng.* **29** 517–26
[19] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk T H and Faubert J 2019 Deep learning-based electroencephalography analysis: a systematic review *J. Neural Eng.* **16** 051001
[20] Waytowich N, Lawhern V J, Garcia J O, Cummings J, Faller J, Sajda P and Vettel J M 2018 Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials *J. Neural Eng.* **15** 066031
[21] Li Y, Xiang J and Kesavadas T 2020 Convolutional correlation analysis for enhancing the performance of SSVEP-based brain-computer interface *IEEE Trans. Neural Syst. Rehabil. Eng.* **28** 2681–90
[22] Ravi A, Beni N H, Manuel J and Jiang N 2020 Comparing user-dependent and user-independent training of CNN for SSVEP BCI *J. Neural Eng.* **17** 026028
[23] Ding W, Shan J, Fang B, Wang C, Sun F and Li X 2021 Filter bank convolutional neural network for short time-window steady-state visual evoked potential classification *IEEE Trans. Neural Syst. Rehabil. Eng.* **29** 2615–24
[24] Zhao X, Du Y and Zhang R 2022 A CNN-based multi-target fast classification method for AR-SSVEP *Comput. Biol. Med.* **141** 105042
[25] Tanaka H, Katura T and Sato H 2013 Task-related component analysis for functional neuroimaging and application to near-infrared spectroscopy data *NeuroImage* **64** 308–27
[26] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces *J. Neural Eng.* **15** 056013
[27] Zhang Y, Cai H, Nie L, Xu P, Zhao S and Guan C 2021 An end-to-end 3D convolutional neural network for decoding attentive mental state *Neural Netw.* **144** 129–37
[28] Borra D, Fantozzi S and Magosso E 2019 Convolutional neural network for a P300 brain-computer interface to improve social attention in autistic spectrum disorder *Mediterranean Conf. on Medical and Biological Engineering and Computing* pp 1837–43
[29] Nakanishi M, Wang Y, Wang Y-T and Jung T-P 2015 A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials *PLoS One* **10** e0140703
[30] Wang H, Zhang Y, Waytowich N R, Krusienski D J, Zhou G, Jin J, Wang X and Cichocki A 2016 Discriminative feature extraction via multivariate linear regression for SSVEP-based BCI *IEEE Trans. Neural Syst. Rehabil. Eng.* **24** 532–41
[31] He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on ImageNet classification *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*
[32] Loshchilov I and Hutter F 2016 SGDR: stochastic gradient descent with warm restarts (arXiv:1608.03983)
[33] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 Rethinking the inception architecture for computer vision *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 2818–26
[34] Zhang C-B, Jiang P-T, Hou Q, Wei Y, Han Q, Li Z and Cheng M-M 2021 Delving deep into label smoothing *IEEE Trans. Image Process.* **30** 5984–96
[35] Müller R, Kornblith S and Hinton G E 2019 When does label smoothing help? *Advances in Neural Information Processing Systems* vol 32
[36] Miyato T, Kataoka T, Koyama M and Yoshida Y 2018 Spectral normalization for generative adversarial networks (arXiv:1802.05957)
[37] Zhang Q, Yang L T, Chen Z and Li P 2018 A survey on deep learning for big data *Inf. Fusion* **42** 146–57
[38] Wei Q, Zhu S, Wang Y, Gao X, Guo H and Wu X 2020 A training data-driven canonical correlation analysis algorithm for designing spatial filters to enhance performance of SSVEP-based BCIs *Int. J. Neural Syst.* **30** 2050020
[39] Li Y, Huang C, Ding L, Li Z, Pan Y and Gao X 2019 Deep learning in bioinformatics: introduction, application and perspective in the big data era *Methods* **166** 4–21
[40] Cheng Y, Wang D, Zhou P and Zhang T 2017 A survey of model compression and acceleration for deep neural networks (arXiv:1710.09282)
[41] Canziani A, Paszke A and Culurciello E 2016 An analysis of deep neural network models for practical applications (arXiv:1605.07678)
[42] Van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605
[43] Yuan P, Chen X, Wang Y, Gao X and Gao S 2015 Enhancing performances of SSVEP-based brain–computer interfaces via exploiting inter-subject information *J. Neural Eng.* **12** 046006
[44] Chiang K-J, Wei C-S, Nakanishi M and Jung T-P 2021 Boosting template-based SSVEP decoding by cross-domain transfer learning *J. Neural Eng.* **18** 016002
[45] Wong C M, Wang Z, Wang B, Lao K F, Rosa A, Xu P, Jung T-P, Chen C P and Wan F 2020 Inter- and intra-subject transfer reduces calibration effort for high-speed SSVEP-based BCIs *IEEE Trans. Neural Syst. Rehabil. Eng.* **28** 2123–35
[46] Wong C M, Wang Z, Rosa A C, Chen C P, Jung T-P, Hu Y and Wan F 2021 Transferring subject-specific knowledge across stimulus frequencies in SSVEP-based BCIs *IEEE Trans. Autom. Sci. Eng.* **18** 552–63
[47] Yan W, Wu Y, Du C and Xu G 2022 Cross-subject spatial filter transfer method for SSVEP-EEG feature recognition *J. Neural Eng.* **19** 036008
[48] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in Neural Information Processing Systems* vol 27
[49] Pan S J, Tsang I W, Kwok J T and Yang Q 2010 Domain adaptation via transfer component analysis *IEEE Trans. Neural Netw.* **22** 199–210
[50] Zhang K, Robinson N, Lee S-W and Guan C 2021 Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network *Neural Netw.* **136** 1–10
[51] Xu L, Xu M, Ma Z, Wang K, Jung T-P and Ming D 2021 Enhancing transfer performance across datasets for brain-computer interfaces using a combination of alignment strategies and adaptive batch normalization *J. Neural Eng.* **18** 0460e5
[52] Krana M, Farmaki C, Pediaditis M and Sakkalis V 2021 SSVEP based wheelchair navigation in outdoor environments *2021 43rd Annual Int. Conf. IEEE Engineering in Medicine & Biology Society (EMBC)* pp 6424–7
[53] Zhang S, Ma K, Yin Y, Ren B and Liu M 2022 A personalized compression method for steady-state visual evoked potential EEG signals *Information* **13** 186