Squeeze and Excitation-Based Multiscale CNN for Classification of Steady-State Visual Evoked Potentials

Jing Jin, *Senior Member IEEE*, Xiao Wu, Ian Daly, Weijie Chen, Xinjie He, Xingyu Wang,

and Andrzej Cichocki, *Life Fellow IEEE*

***Abstract*—Brain-Computer interface (BCI) technology enables the control of external devices by recognizing user intentions. Steady-state visual evoked potential (SSVEP)-based BCI technology has been widely applied in the field of Internet of things (IoT) device control, including smart healthcare, smart homes, and robotics, and has achieved significant results. However, as the field of BCI-based IoT device control is still in its development stage, there remains considerable room for improvement in terms of accuracy, efficiency, and cost. Therefore, enhancing the classification accuracy of SSVEP decoding using a short time window, reducing both human and material costs, and improving work efficiency are crucial for the theoretical research and engineering applications of BCI technology in IoT device control. Based on this, we propose a novel approach to address the challenge of high accuracy feature extraction within brief timeframes. Our approach integrates a multi-scale convolutional neural network with a squeeze excitation module (SEMSCNN). This fusion leverages CNNs' local feature learning capacity and the advantageous feature importance distinction offered by the squeeze excitation mechanism. First, the EEG signals are band-pass filtered into distinct frequency bands and frequency band and channel features are extracted by a two-layer convolution. Then, temporal features are extracted via a multi-branch convolution of different scales. Finally, the squeeze and excitation (SE) module is introduced to learn the interdependence between features to improve the quality of the extracted features. The first stage of training exploits statistical commonalities across research participants by learning the global model, and the second stage fine-tunes each participant's features separately by exploiting participant-specific differences in features. We evaluate our SEMSCNN model on two large public datasets, Benchmark and BETA, and we compare our model to other state-of-the-art models in order to evaluate the effectiveness of our proposed network. Our experimental results indicate that our method effectively improves the accuracy of target recognition and information transfer rate under short-duration stimuli, showing a significant advantage compared to other baseline methods. This provides a broad prospect for the practical application of BCIs in the field of IoT.**

***Index Terms*—Multiscale fusion, convolutional neural network (CNN), squeeze and excitation module (SEM), Brain–computer interface (BCI), steady-state visual evoked potentials (SSVEP)**

I. INTRODUCTION

Brain-computer interfacing (BCI) is a technology that enables a connection between the brain and external devices [1-4]. In recent years, Brain-Computer interfaces (BCIs) have become increasingly popular in fields such as IoT device control, entertainment, and communication [5]. Electroencephalogram (EEG) signal components that are commonly used to control BCIs include event-related potentials (ERPs) [6-10], steady-state visual evoked potentials (SSVEPs) [11-14], slow cortical potentials (SCPs) [15], and sensorimotor rhythms [16-19]. Among them, SSVEP has become an important set of control signals due to its high signal-to-noise ratio (SNR) [20-23]. This characteristic allows SSVEP-based BCI technology to achieve high-precision decoding, high information transfer rates, and short window length decoding [24] . Particularly in the context of IoT device control, users can generate SSVEP signals by focusing on specific visual stimuli. These signals can be decoded into control commands to operate devices such as lights, air conditioning, and music players. This contactless control method significantly enhances convenience

This work was supported by the Grant National Natural Science Foundation of China under Grant 62176090 and STI 2030-major projects 2022ZD0208900; in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX, This research is also supported by Project of Jiangsu Province Science and Technology Plan Special Fund in 2022 (Key research and development plan industry foresight, fundamental research fund for the central universities JKH01231636 and key core technologies) under Grant BE2022064-1.

Jing Jin are with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China, and also school of Mathematics, East China University of Science and Technology, Shanghai 200237, China (e-mail: jinjingat@gmail.com;) (Corresponding author: Jing Jin).

Xiao Wu, Weijie Chen, Xinjie He, and Xingyu Wang are with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237,China(email:wuxiao121409@163.com;wjchen827@foxmail.com;xinjieHe@mail.ecust.edu.cn;xywang@ecust.edu.cn).

Ian Daly is with the Brain-Computer Interfacing and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK (e-mail: i.daly@essex.ac.uk).

Andrzej Cichocki is with the Systems Research Institute, Polish Academ of Science, 01-447 Warsaw, Poland, and with RIKEN Advanced Intelligence Project, Tokyo 103-0027, and also Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan (e-mail: a.cichocki@riken.jp).

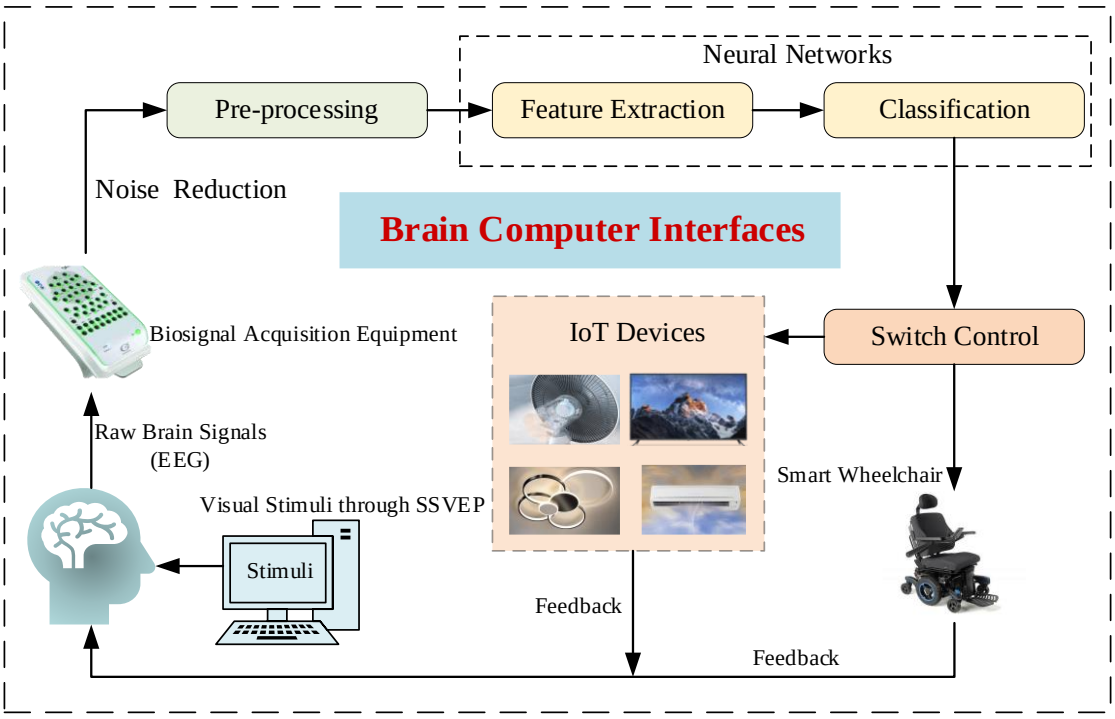


Fig.1. Flowchart for controlling external devices using Brain-Computer Interface

and accessibility, especially for users with limited mobility (as shown in **Fig.1**). Furthermore, when integrated with IoT technology, SSVEP-based BCI can create smarter home environments, greatly improving the quality of life for users.

A BCI system based on SSVEP signals attempts to detect changes in EEG signals in the occipital visual area that are informative about the stimulus the BCI user is attending to. When we experience a visual stimulus that oscillates at a fixed frequency, the potential activity within our cerebral cortex will be modulated so as to produce a continuous response which has a periodic rhythm similar to the stimulus frequency (the fundamental frequency or the harmonic frequency of the stimulus frequency). This periodic response to a regular stimulus is referred to as the SSVEP. In the SSVEP BCI control paradigm, the BCI user is asked to fixate on a flashing target while the EEG is recorded. To recognize the intention of the BCI user, the collected EEG signals are analyzed to identify which target frequency they relate to most strongly.

Different recognition methods have been attempted to improve decoding accuracy of SSVEP-based BCIs. Canonical correlation analysis (CCA) is often used to perform frequency identification of SSVEPs [25-29]. Specifically, recognition is performed by calculating the correlation between the collected EEG signals and template signals of the stimulus target. However, the recognition performance of this method is poor when using short time windows. To boost the SSVEP recognition accuracy for short time windows, an improved CCA method called CCA-M3 was proposed [26]. In traditional CCA, there is no significant difference in the classification accuracy of the reference signals with different numbers of harmonic frequencies [25]. Therefore, some researchers have proposed a filter bank CCA (FBCCA) method, which combines the fundamental frequency and harmonic frequency components to improve the classification performance [24, 30, 31].

The correlation component Analysis (CORRCA) algorithm maximizes the correlation between the multi-channel template signal (calculated by averaging SSVEP signals from multiple trials for each frequency in the training set) and the multi-channel test signal. The frequency with the highest correlation then indicates the final recognition target [30]. Maximization in CORRCA [30] is a single projection across channels, while maximization in standard-CCA [27] is two projections, one of which is across channels and the other uses harmonics in the reference. An extension of CORRCA, filter-bank CORRCA (FBCORRCA), uses a filter-bank approach [30], while hierarchical feature CORRCA (HFCORRCA) uses exponentially decaying weights to fuse information from other correlation coefficients [32], and two stage CORRCA (TSCORRCA) uses spatial filters for all stimulus frequencies to form a better performing extension [30]. In contrast to the traditional CCA method, in order to extract and enhance the EEG components most relevant to a specific visual stimulation task, a method called task-related component analysis may be used for stimulus target recognition of SSVEPs [24]. This method uses the participant's own EEG signal as a template to effectively extract task-relevant components by maximizing reproducibility during the task. However, the stability of task related component analysis (TRCA) is easily affected by the participants. In order to improve the generalization ability of TRCA, an ensemble learning framework, eTRCA, was introduced to provide a more accurate and stable processing of SSVEP signals [24]. eTRCA makes up for the disadvantage of TRCA's weak adaptability, and further improves the

recognition performance of SSVEPs.

Over the last few years, deep learning (DL) methods have been gradually applied to stimulus target recognition of SSVEPs. For example, the convolutional neural network (CNN) method, a popular image classification method, may be used to learn the local key features of an image. This better learning ability provides a blueprint for analyzing EEG signals using CNN. Although there are many examples of SSVEP classification achieving impressive performance [33-35], further advancements in accuracy are still possible.

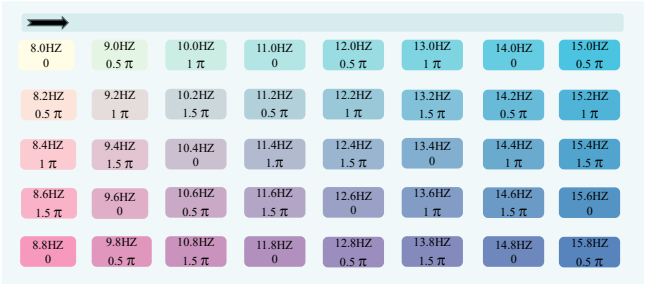
CNN are one of the biggest advances in machine learning in recent years [36-39]. CNNs allow for automatic feature extraction within their layers, which means the model is capable of learning features without the need for manual feature extraction. In addition, SSVEPs have rich internal information. CNNs have advantages over traditional machine learning models in dealing with such internal information including good fault tolerance, self-learning ability, good adaptive performance, and high resolution.

CNNs have been used to decode brain signals in some recent studies [40-42]. For example, Cecotti proposed a CNN architecture that simulates the FFT process to classify SSVEP signals in the time domain [34]. In 2017, Kwak et al. proposed a deep classifier based CNN that uses frequency features as inputs for robust SSVEP detection and found very encouraging SSVEP decoding results [35]. However, most CNNs rely on the frequency domain features after performing a Fast Fourier Transform (FFT) as the inputs to the network. These features have obvious long time window characteristics [43, 44]. Considering that methods such as FT have poor ability to extract frequency features for signals with short time windows, rich time domain information will also be lost. To resolve this issue, Guney et al. proposed a novel BCI SSVEP speller DNN architecture, which processed SSVEP signals in the time domain as an end-to-end system from EEG acquisition to target feature prediction [45].

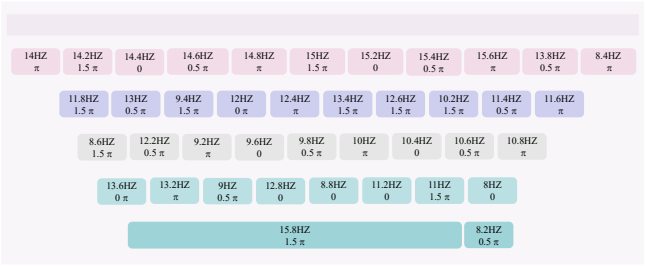
Considering the singularity of the network structure of the DNN architecture and the imperfection of feature extraction, we propose a new network architecture called SEMSCNN to classify SSVEPs. First, the original EEG is band-pass filtered to obtain different frequency bands of EEG and the frequency and channel features are extracted by two-layer convolution. Then the temporal features are extracted by multi-branch convolution at different scales and the features of different branches are fused. Finally, the squeeze and excitation (SE) module is introduced to learn the interdependence between features to improve the quality of the extracted features. The contributions of our proposed work are summarized as follows.

- First, in contrast to the single-branch network architecture, our proposed SEMSCNN architecture contains three convolution kernels of different scales. This approach aims to improve the extracted time domain information while correcting the issues of insufficient feature extraction using single branches.
- Second, we introduce an SE module after the feature merging over different scales. We aim to improve the quality of feature extraction by learning the interdependence between features through the SE module.
- Finally, our proposed SEMSCNN architecture achieves ITRs of 204.55 bits/min and 141.18 bits/min with only 0.2 seconds of stimulus time on the Benchmark and BETA datasets. The ITR is higher than the performance results reported by DNN on these datasets. The high decoding accuracy lays a solid foundation for the practical application of BCI technology in the field of the IoT.

The rest of this article is organized as follows. Section II presents a detailed introduction to our proposed method. To evaluate the effectiveness of our proposed network, our experimental results are showed in Section III. Section IV and V then report the discussion and conclusion.



(A)



(B)

Fig.2. (A) Stimulus target distribution on the Benchmark dataset, (B) Stimulus target distribution on the BETA dataset.

II. MATERIALS AND METHODS

We first introduce the datasets we are using and then describe our proposed SESCNN architecture and its main design components.

A. Public Dataset

The first dataset we used is the Benchmark dataset. This dataset is obtained from a SSVEP-BCI speller with 40 targets. Joint frequency-phase modulation (JFPM) was used to encode all stimulus flicker frequencies and stimulation frequencies ranged from 8 Hz to 15.8 Hz with 0.2 Hz intervals. The phase ranged from 0 to 1.5π with an interval of 0.5π. A 64-electrode EEG recorder was employed for acquisition, sampling EEG at 250 Hz. The dataset consists of EEG recordings from 35 healthy participants. For each participant, the data consisted of 6 blocks of 40 trials corresponding to all 40 targets displayed by visual

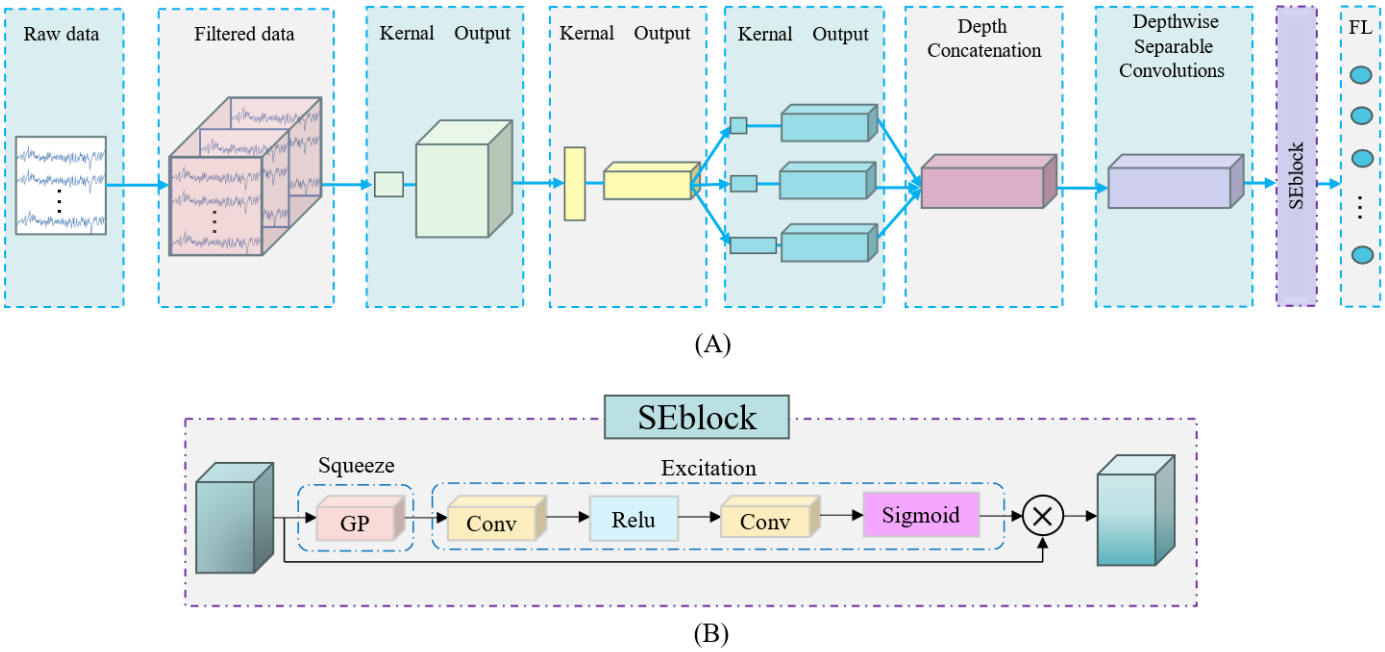


Fig. 3. (A) Detailed framework of SEMSCNN, (B) Detailed network of SEblock

cues in random order. The target stimulus distribution is shown in Fig. 2(A). Each trial comprised 0.5 seconds of EEG recording both before and after the 5-second stimulus. For more detailed descriptions of the data acquisition, please refer to [46].

The second dataset we used is the BETA dataset. The BETA dataset is similar to the Benchmark dataset, but there are some differences. A total of 70 healthy participants are recorded in this dataset. Each experiment consisted of four blocks. A blinking target character is displayed in the form of a keyboard (see **Fig. 2(B)**). The experiments were performed outside the laboratory environment and the SNR is, consequently, low compared to the Benchmark dataset. Therefore, object recognition is more challenging in this dataset. The stimulus duration is 2 seconds for the first 15 participants and 3 seconds for the remainder of the participants. The average visual delay of the participants in this dataset is about 130ms. For a detailed description of the data, please refer to [47]. To minimize computational complexity, the collected data are filtered to improve the feature learning ability of our network. The EEG signals within the frequency bands 8-90 Hz, 16-90 Hz, and 24-90 Hz are then obtained by the Butterworth bandpass filter.

B. Proposed SEMSCNN Architecture

Our SEMSCNN is an end-to-end system designed to process multichannel raw EEG signals. Instead of relying on manually extracted features, the method learns and extracts features based on the properties of the raw data, ultimately realizing the multi-target recognition task of SSVEP identification. First, the network extracts band and channel information through two convolution layers, it then extracts different time domain features through three parallel multi-scale convolutions. Then, the time domain features over different scales are fused and depth separable convolution is used to extract the depth information from the fused time domain features. An SE module is then introduced to enhance the network's ability to represent the input data and realize feature recalibration. This allows the network to enhance important features and suppress unimportant features. Finally, the multi-stimulus target classification of SSVEPs is realized by the fully connected layer, the activation layer, and the classification layer. The entire framework of our network is shown in **Fig. 3(A)**. The network elements are explained in detail below.

The first layer for frequency band feature extraction: The contribution of different harmonics in the SSVEP signal may vary with the frequency of the flicker stimulation of the target signal. In general, the lower harmonics have a higher amplitude, while the higher harmonics, although fewer in number, often show a higher signal-to-noise ratio because they interfere less with other ongoing activities in the brain due to the 1/f frequency power distribution of the EEG [48] [49]. However, it is not trivial to decide which harmonic is more informative when dealing with multiple sub-band signals of the SSVEP separately or when using restricted models for signal fusion [48] [50]. However, the optimal method to accurately set the weight of each harmonic has not been fully explored in the literature.

In our SEMSCNN design, we choose to remain agnostic about the normalization of harmonics and instead let the network learn, in a data-driven manner, the normalization weights. Therefore, we perform bandpass filtering on multi-channel SSVEP signals. For each channel, the lowest cutoff frequency of the SSVEP signal $x \in R^{C \times N}$ is $r \times \min\{f_j\} - \epsilon HZ$ (eg. 8 HZ for r=1) and the highest cutoff frequency is $6 \times \max\{f_j\} + \epsilon HZ$ (eg. 90 HZ). Here, ϵ is the small margin. A zero-phase Chebyshev type 1 filter with filter order 2 and passband ripple of 1 dB is designed using the MATLAB design

filter function. Thus, each filter excludes harmonics of order less than r and includes the remaining harmonics up to order 6 (the maximum order is set to 6, since in EEG frequency components over 100 HZ is usually considered to be noise).

The first layer of our SEMSCNN network learns the weights $w \in R^{Ns\times 1}$ of the subbands, linearly combining the subbands for normalization across harmonics, such that $z = [x^{(1)},\dots,x^{(r)},\dots,x^{(Ns)}]w$, where the input to the layer is $[x(1),\dots]\,x(r),\dots,x(Ns)] \in R^{C\times N\times Ns}$ (i.e. for $C = 9, N = 50 = T \times fs, T = 0.2s, fs = 250HZ, Ns = 3$, the size is $9 \times 50 \times 3$), the size of the convolution kernel is (1,1), and the output $z1 \in R^{C} \times N$ (i.e. for C = 9, N = 50, the size is $9 \times 50 \times 1$). Therefore, our SESCNN network can learn the weights according to the properties of the data itself and extract effective frequency band information

***The second layer for the extraction of channel information*:** In the process of collecting EEG containing SSVEP signals, multiple electrodes are typically used to capture brain activity from multiple scalp locations. This method of signal capture allows researchers to analyze brain responses to visual stimuli from multiple perspectives. Multi-channel data provides information about the spatial distribution of electrical activity in the brain, which helps to localize the signal source more accurately. Within the visual cortex different regions may have different responses to specific frequencies of visual stimuli and, by analyzing the activity of these regions, SSVEP signals can be better understood and utilized. In our neural network architecture for processing these signals the second layer of the network plays the crucial role of extracting information from each channel. In this layer, it is common practice to use a convolution operation, which can extract features from each channel efficiently. The convolution operation processes the input multi-channel EEG signals by applying multiple filters, each designed to capture specific signal characteristics, such as the spatial pattern or frequency distribution of the signal. Consequently, the convolutional layer can integrate information from various electrodes, leading to a more complete understanding and utilization of the complex SSVEP-related information found in EEG signals.

The size of the convolution kernel in the second layer of our network is (9,1) and the number of output neurons is 120. After extracting the frequency band information, the output size of the data is $9 \times 50 \times 1$ and this data is used as the input to the channel convolution layer. After the channel convolution, the output size of the data is $1 \times 50 \times 1$ (when using the above data as an example, with a time window of 0.2 seconds). We hypothesize that the information between channels is crucial to improve the performance of SSVEP target recognition.

Multi-scale convolutions for temporal domain feature extraction : In the visual cortex of the brain, SSVEPs are produced in response to visual stimuli presented at designated frequencies. These SSVEP responses may behave differently at different time scales. Multi-scale convolution can be used to capture these feature changes over time to distinguish stimuli with different frequency distributions more effectively. To this end, we introduce a multi-scale convolution layer. By using filters of different sizes to analyze the signals, the time domain information can be extracted at different time scales. We introduce three parallel convolution branches into our SEMSCNN network architecture, the convolution kernel size is (1,3), (1,9), (1,13) and the number of output features is 120.

Our network architecture directly processes the raw electrical signal and is an end-to-end system. Multi-scale convolution not only reduces the noise in EEG signals but also effectively extracts useful signal features. This is particularly advantageous in situations involving weak signals or significant background noise, as it helps retain key information more effectively.

Due to the large inter-person differences in SSVEP signals, multi-scale convolution may be suitable to adapt to the specific signal characteristics of different users by adjusting filters at different scales, improving the generalization ability and flexibility of the system. Using multi-scale convolution can, therefore, potentially improve the recognition performance of SSVEPs over multiple participants.

Concatenation, depth wise separable convolution and SE modules for feature enhancement: After multi-scale convolution, features at different scales usually capture different information. When processing SSVEPs, larger-scale features may capture more global and stable signal properties, while smaller-scale features may be more sensitive to rapidly changing signal details. To obtain a richer and more comprehensive signal representation it is necessary to concatenate the features over different scales. Therefore, a concatenation layer is introduced. Comprehensive use of multi-scale features can help the model better distinguish the target signal from the background noise, which is helpful to improve the performance of our SEMSCNN model.

To further extract deep temporal features, we introduce depth-wise separable convolution. Depth-wise Separable Convolution is an efficient convolutional neural network architecture originally proposed by Chollet [51]. This technology achieves the purpose of reducing computational cost, reducing model size and improving operating efficiency by reconstructing the traditional convolution. Specifically, depth-wise separable convolution consists of two steps: depth wise convolution and pointwise convolution (1×1 convolution). When performing deep convolution, if the number of channels in the input feature map is N, a convolution kernel is used for each of the N channels to obtain N feature maps with 1 channel each. Then the N feature maps are concatenated in order to obtain an output feature map with N channels. Finally, a 1×1 convolution operation is performed to fuse different channels.

In the traditional CNN architecture, convolutional and pooling layers are usually used to extract features. However, this approach does not explicitly model the relationship between feature channels, resulting in some channels contributing relatively little to a specific task while others are more important. The SE module aims to solve this problem. In the realm of DL, the SE module is recognized as a promising attention mechanism with superior adaptive feature extraction performance. Proposed by Jie Hu et al. in 2017, it aims to

Table I
Parameters of each layer of our SESCNN network (L=time×sampling frequency)

Layer type	Output dimension	Kernel size	Step size	Output Shape	Options
Input				(9, L,3)	
Conv2d	1	(1,1)	1	(9, L,1)	
Conv2d	120	(9,1)	1	(1, L, 1)	
Dropout					Ratio=0.1
Conv2d	120	(1,3)	1	(1, L, 1)	Mode=same
Conv2d	120	(1,9)	1	(1, L, 1)	Mode=same
Conv2d	120	(1,13)	1	(1, L, 1)	Mode=same
Concatenate	360			(1, L, 1)	
DpwConv2d	120	(1,3)	1	(1, L, 1)	
Dropout					Ratio=0.95
SE	120			(1, L, 1)	
Dense	40				
Activation					softmax

improve the efficiency of information transfer between CNN channels [52]. The SE module models the relationship between channels by introducing a Squeeze operation and an Excitation operation.

In the Squeeze stage, the module compresses the output feature map of the convolutional layer into a feature vector via a global average pooling operation (e.g. the output feature map is of size $W \times H \times C$, which becomes $1 \times 1 \times C$ after squeezing). Then, in the Excitation stage, a fully connected layer and a nonlinear activation function are used to learn to generate a vector of weights for a channel (still of data size $1 \times 1 \times C$). This weight vector is applied to each channel on the original feature map to weight the features of the different channels (the weights are multiplied with the original feature map and the data size is restored to $W \times H \times C$). In this way, the SE module is able to adaptively learn the importance of each channel, and the channel contributions in the feature map are weighted according to the needs of the task. By learning the interdependence between features, the quality of the extracted features is improved. In our proposed SEMSCNN network, all fully-connected layers in the conventional SE module are replaced by 1-D convolutional layers with kernel 1×1. The specific structure of our SEMSCNN network is shown in **Fig.3 (B).**

After the temporal feature extraction, the final classification results are output by the Final Layer (FL). In **Fig. 3 (A)**, the FL specific package contains three layers, the first layer is a fully connected layer, the second layer is a sigmoid activation layer, and classification occurs in the third layer. The parameters of the network are shown in Table I (note: DpwConv2d is shorthand for depth wise separable convolution).

C. Two-Staged Training of SEMSCNN

When attempting SSVEP target recognition, training an effective neural network model is crucial for improving recognition accuracy. To ensure the model captures a broad range of data features and trends a two-stage training method is employed. First, the initial phase involves global training using the entire training set (the training method in the first phase is leave-one-block out cross-validation), which provides a global perspective that helps ensure the model's generalizability and robustness across different SSVEP tasks. However, each SSVEP target may be associated with specific frequencies or other conditions, necessitating a more detailed understanding and optimization of the model for these specific scenarios.

Therefore, the second stage, the participant-specific fine-tuning stage, seeks to further enhance the responsiveness of the model to each specific target or frequency. In this stage, the model is trained for each participant using only the subset of data from that participant taken from the global model parameters obtained in the first stage. This targeted training enables the model to better adjust and optimize the processing of each specific target, improving the recognition accuracy and recognition speed for each participant.

In stage 1 “Global training” the full training set is used to construct a base architecture with a broad field of view, which is able to capture the common characteristics of various SSVEP signals, thus laying the foundation for DL of specific targets. Training the model using all available training data over all participants, helps the model find universally applicable patterns across different types of input data.

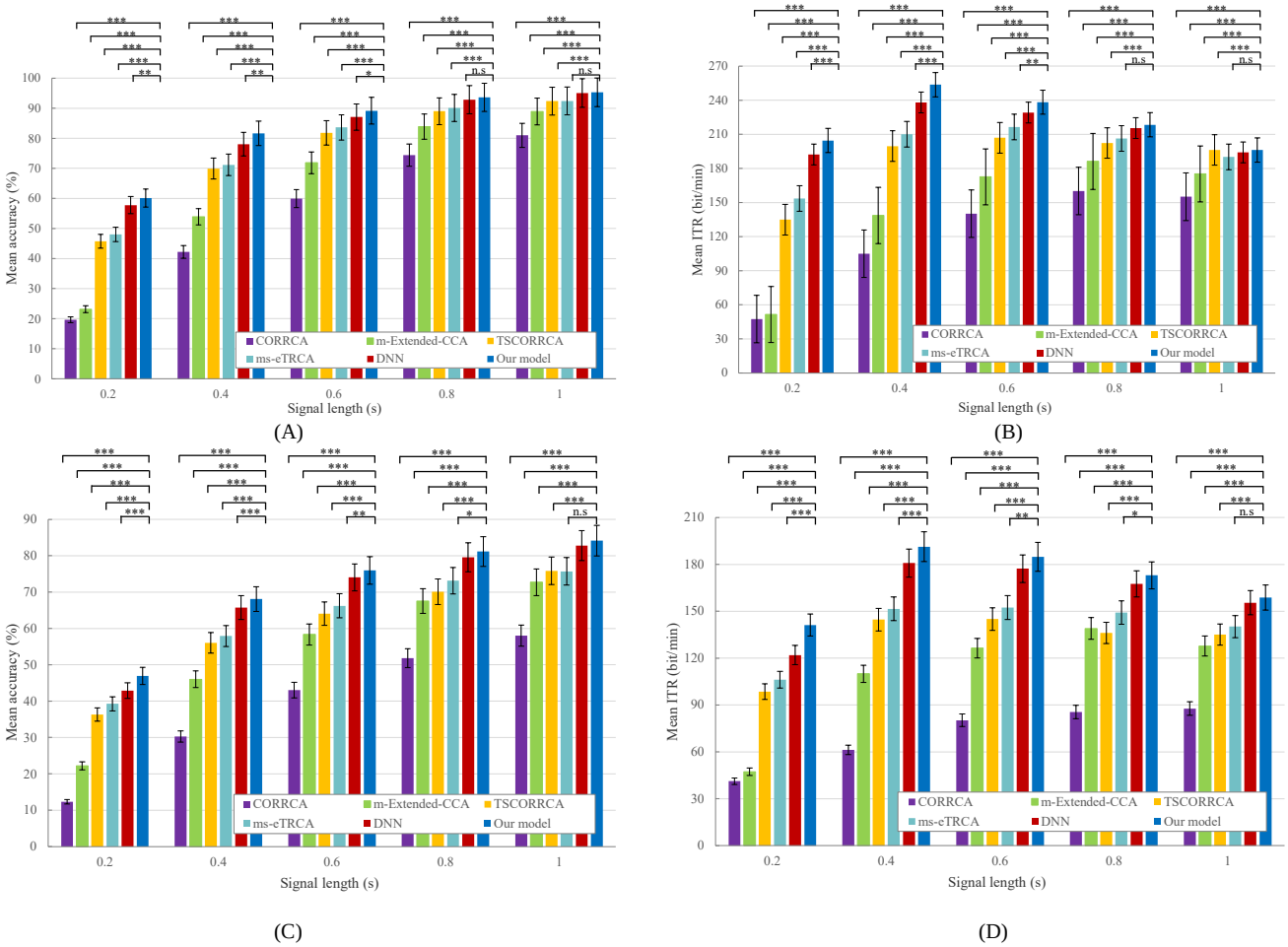


Fig.4. (A) Classification accuracies achieved on the Benchmark dataset. (B) ITRs achieved on the Benchmark dataset. (C) Classification accuracies achieved on the BETA dataset. (D) ITRs achieved on the BETA dataset. The single asterisk * denotes p<0.05; ** denotes p<0.01; and ***denotes p<0.001

In stage 2 "Participant-specific Fine-tuning" the model is refined and specialized according to the specific requirements of the SSVEP task (e.g., specific frequency identification). From the model parameters obtained by global training, a subset of data relevant to a particular participant is selected to retrain or fine-tune the model. This targeted optimization can make the model perform better under certain conditions. Specifically, in each iteration, we train the network based on the training batch data $\{(x_i,\ y_i)\}_{i=1}^{D_b}$,where D_b is the number of trials in the batch, by minimizing the categorical cross-entropy loss

$$\frac{1}{D_b}\sum_{i=1}^{D_b} -\log\left(s_i(y_i)\right) + \lambda|w|^2$$

via the Adam optimizer [53] with a learning rate $\nu = 0.0001$ (without decaying), where λ is the constant of the L2 regularization, which we set as $\lambda = 0.001$, $s_i \in [0,1]^{M\times 1}$ is the softmax output for the instance x_i, $s_i(y_i)$ is the y_i'th entry of s_i and the final prediction is $\hat{y}_i = argmax s_i(j)$. Here, w represents all the SEMSCNN weights.

This two-stage training strategy is very valuable in SSVEP target recognition, because it combines the advantages of global learning and local fine adjustment, which can not only ensure the generalization ability of the model, but also meet the needs of specific tasks. In this way, the performance and accuracy of the model can be effectively improved.

The hardware information of the computer used in our experiments is as follows: 11th Gen Intel(R) Core (TM) Intel(R) Core (TM) i9-14900HX @ 2.20GHz, NVIDIA GeForce RTX 4080 Laptop GPU.

D. Comparison Algorithms

We compare our proposed model to the following state-of-the-art models for SSVEP decoding.

1) CORRCA [8]: The final prediction is generated by maximizing the correlation between the multichannel template signals, computed by averaging the SSVEP signal across multiple trials for each frequency in the training set, and the multichannel test signal, and then selecting the frequency with the highest correlation.
2) m-Extended-CCA [54]: An extension based on the CCA method to improve the classification performance of SSVEP target recognition.
3) TSCORRCA [8]: A spatial filter over all stimulus frequencies is utilized to produce more discriminative features for SSVEP target frequency identification.

4) ms-eTRCA [55]: A cross-multiple stimulus learning scheme is used for frequency stimulus target recognition. This scheme is suitable not only for learning data corresponding to the target stimulus, but also for learning data corresponding to other stimuli.
5) DNN [15]: DNN processes the multi-channel SSVEP with convolutions across the sub-bands of harmonics, channels, and time. It then performs classification in the fully connected layer.

Table II
Accuracy (%) and standard deviation (%) with 0.4s time length

Dataset / Channels	Benchmark	BETA
3 channels	53.21±2.98	43.29±3.54
6 channels	77.21±3.01	45.43±3.20
9 channels	81.65±3.31	68.07±3.47
32 channels	82.24±3.29	66.30±3.57

E. Performance Evaluations

In order to verify the effectiveness of our proposed method, we conducted experiments on two public datasets: Benchmark and Beta. We compare our model to five methods: CORRCA, Extended-CCA, TSCORRCA, ms-eTRCA, and DNN. In our comparisons, the same test procedure for all these methods is followed. Accuracy and Information Transfer Rate (ITR) are used to evaluate the target recognition system.

The ITR is a measure of system efficiency, which takes into account not only the accuracy, but also the recognition speed, and number of conditions. ITR is measured in bits per minute (bpm) and reflects how much information can be transferred per second by the BCI. A high ITR value implies not only a high accuracy but also a fast response, which is particularly important for BCI application environments. The ITR in bits per minute (bpm) is calculated by[42]:

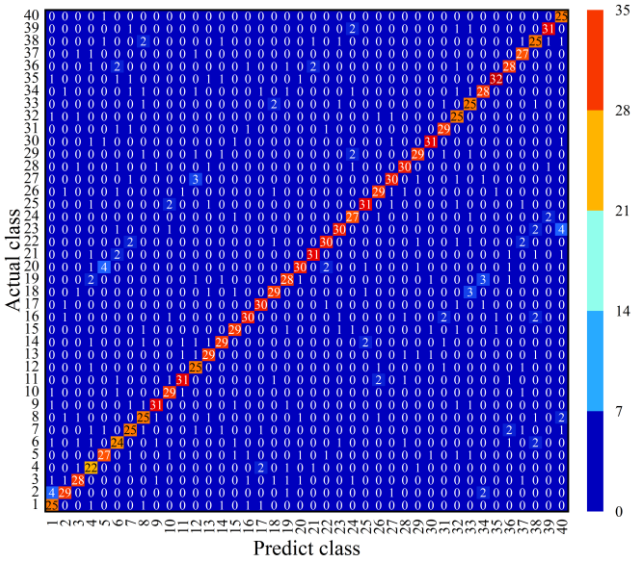
$$ITR = \frac{60}{T_W}\left[log_2 N_f + PLog_2 P + (1-P)log_2\frac{1-P}{N_f-1}\right]$$

where T_W is the time length of the test signal, N_f is the number of stimulus targets (N_f=40 on our two public datasets), and P is the classification accuracy. The statistically significant difference between the two conditions is determined by employing the paired *t*-test.

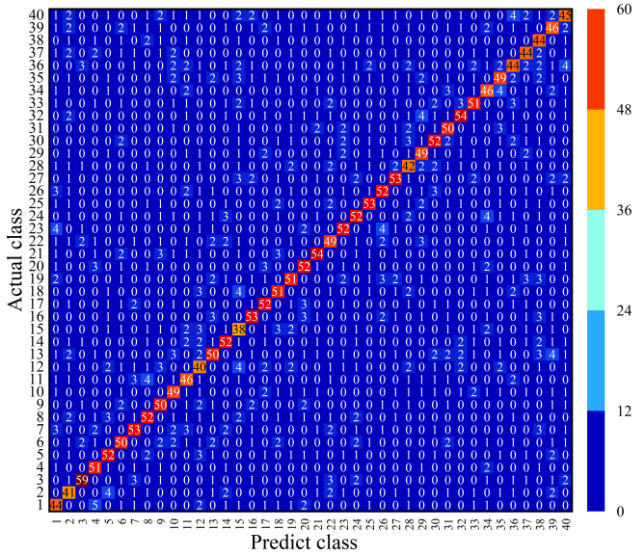
III. RESULTS

A. Performance Evaluations with the Benchmark dataset

A total of 35 participants are included in the Benchmark dataset and each completed 6 blocks with 40 target SSVEPs. EEG signals with three frequency bands and nine channels were used as input to the SEMSCNN network for multi-stimulus target classification. The time window lengths used are 0.2s, 0.4s, 0.6s, 0.8s, and 1s. The average accuracy over all participants and time windows is shown in **Fig. 4(A)**.



(A)



(B)

Fig.5. (A) Confusion matrix of SEMSCNN based on the Benchmark dataset. (B) Confusion matrix of SEMSCNN based on the BETA dataset.

Specifically, in a short time window of 0.2s, the average accuracy over all participants is as high as 60.10%. With the increase in the length of the time window, the accuracy increases: 81.65% (0.4s), 89.17% (0.6s), 93.6% (0.8s) and 95.64% (1s). **Fig.5(A)** shows the stimulus confusion matrix with 9 channels and 0.4 second windows. The diagonal lines identify the correct number of classifications for each target. As the length of the time window increases, the accuracy increases as the extracted features become richer due to the increasing amount of data. In the time window of 0.2s, the accuracy of our proposed method is 2.34% higher than the accuracy achieved with DNN and 12% higher than the accuracy achieved with ms-eTRCA. Our proposed method shows a highly significant ($p < 0.001$) improvement in accuracy compared with CORRCA, m-

TABLE III
Ablation studies of SESCNN on Benchmark dataset

Time windows	With SE	Without SE	Before SE
0.2s	60.10 ±3.22	58.97±3.02	54.26±3.48
0.4s	81.65±3.31	79.93±3.53	77.77±3.44
0.6s	88.64±2.65	88.08±2.98	86.54±3.10
0.8s	93.60±2.74	93.25±3.23	92.30±2.65
1.0s	95.64±2.72	95.20±2.70	94.40±2.95

TABLE IV
Ablation studies of SESCNN on Beta dataset

Time windows	With SE	Without SE	Before SE
0.2s	46.94±3.12	45.19±3.12	36.99±3.09
0.4s	68.07±3.47	63.08±4.29	59.29±3.72
0.6s	75.96±4.13	71.93±4.15	68.67±4.47
0.8s	81.14±4.31	78.23±4.67	75.14±4.34
1.0s	84.13±4.69	81.37±4.44	78.89±4.66

TABLE V
Computational complexity comparison among the proposed models and baseline models

	Model	Parameters	Train time(s)	Test time(ms)	Energy consumption (KJ)
Benchmark dataset	DNN	654.284K	5827.32	22.31	406.75
	MSCNN	2.201M	13470.39	55.92	1027.79
	SEMSCNN	2.215M	17228.73	27.59	1323.17
BETA dataset	DNN	653.080K	8222.98	22.87	279.58
	MSCNN	2.199M	21887.16	54.57	1703.48
	SEMSCNN	2.211M	24570.87	64.44	1904.24

Extended-CCA, TSCORRCA, ms-eTRCA, and a significant ($p < 0.01$) improvement compared with DNN. In addition, in the time windows of 0.4s and 0.6s, compared with the other five methods, SEMSCNN achieves significantly better accuracies. When we compare our method with DNN in the 0.8s and 1s time windows, although there is no significant difference, the accuracy is approximately 1% higher.

The ITR in different time windows is shown in **Fig.4(B)**. For the average ITR over 35 participants, our proposed method can reach up to 256.33bits/min, which is 15.6 bits/min higher than the ITR of DNN, within a time window of 0.4s. The ITR of our proposed method reaches 204.55 bits/min with a short time window of 0.2 seconds. Under the time length of 0.2s, 0.4s, 0.6s, compared with the other five comparison methods, our method is significantly better.

B. Performance Evaluations of the Beta dataset

There are 70 participants in the Beta dataset and the results presented are the average over all participants. Each participant completed four blocks in which they selected among 40 stimulus targets. As with the Benchmark dataset, three frequency bands and nine channels are used as input to our SESCNN network. The average classification results over all participants and time windows are shown in **Fig.4(C)**. **Fig.5(B)** shows the stimulus confusion matrix with 9 channels and a 0.4 second window. The diagonal lines identify the correct number of participants for each target. First, within a short time window of 0.2s, the accuracy of our method is 46.94%, which is 34.63%, 24.71%, 10.62%, 7.68%, and 4.05% higher than that of CORRCA, m-Extended-CCA, TSCORRCA, ms-eTRCA, and DNN methods respectively. Except for the time length of 1s, which is not significantly different to DNN, our method is significantly better than the other comparison methods at all time window lengths.

In terms of ITR, our proposed method also performs very well. The ITR over different time windows is shown in **Fig.4(D)**. The average ITR over all 70 participants can reach up to 191.32 bits/min. This is 10.47 bits/min higher than the highest ITR achieved with DNN. When using window lengths of 0.2s and 0.4s, there are extremely significant differences between our method and CORRCA, m-Extended-CCA, TSCORRCA, ms-eTRCA, and SESCNN, as well as significant differences between our method and DNN and SESCNN, at a window length of 0.8s.

IV. DISCUSSION

A. The influence of the number of channels on SEMSCNN

In BCI systems, multiple electrodes are usually used for data acquisition. Different electrode channels may have different signal quality and noise levels, therefore, selecting channels with high signal quality and low noise can significantly improve the accuracy of SSVEP detection. In addition, reducing the

number of channels for data processing can reduce the computational cost and improve the processing speed, which is particularly important for real-time systems. Channel selection also improves device portability and user comfort, as using fewer electrodes makes the device more portable and easier to configure. Finally, since brain structure and electrophysiological properties may vary among individuals, channel selection can also optimize the electrode configuration according to individual differences to provide the best signal capture for each user. By carefully selecting channels the overall performance and user experience of the SSVEP system can be improved. Therefore, we selected data from different channels to validate our proposed SESCNN model. **Table II** reports the accuracy with 3 (O1, Oz, O2), 6 (O1, Oz, O2, POz, PO3, PO4), and 9 (Pz, PO3, PO5, PO4, PO6, POz, O1, Oz, O2) channels as typically used in the literature, and 32 channels (all channels from central-parietal regions and C3, C1, Cz, C2, C4, FCz). In the table we report the accuracy and standard deviations when using the 0.4s time window for both datasets.

According to **Table II,** the accuracy is the lowest when using data from only three channels. As the number of channels increases, the accuracy gets higher and higher. In particular, it can be seen that the accuracy is 81.65% when the number of channels is 9 and 82.24% when the number of channels is 32. The two results are not very different, but the number of channels is very different. The greater the number of channels, the larger the amount of data, the higher the computational complexity needed to train the model, and the longer the training time of the model. Therefore, it is a better strategy to select 9 channels. Under the condition of ensuring high performance, fewer channels are more practical for SSVEP-based BCIs.

B. Ablation Studies

Our proposed SEMSCNN network model chiefly consists of three parallel multi-scale convolution and SE modules. In order to verify the effectiveness of the introduction of the SE module, we conducted ablation experiments on our two datasets. Specifically, we conducted studies with and without SE modules. On the benchmark dataset, when the window length is 0.6s, 0.8s, and 1s, there is no significant improvement as a result of including the SE module. However, an improvement in the performance of the model is observed when short time windows of 0.2s and 0.4s are used (see **Table III** for detailed results).

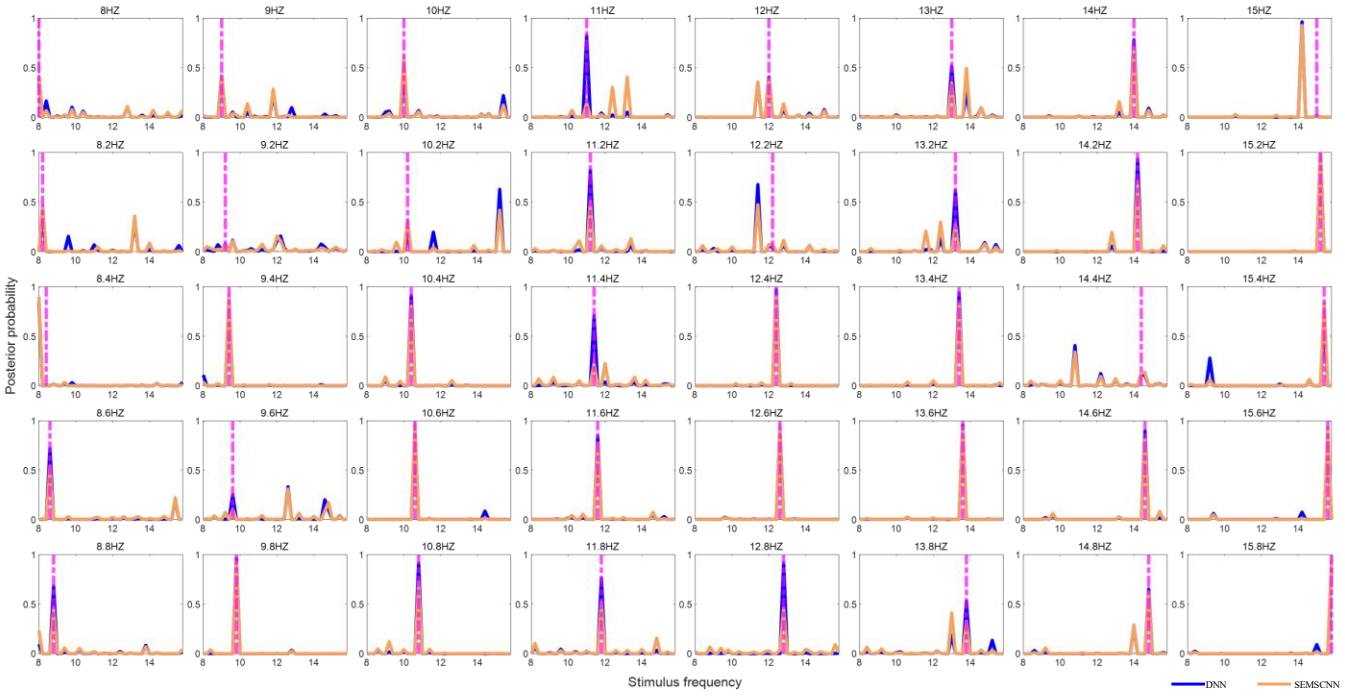
The classification results for the same ablation study on the BETA dataset are shown in **Table IV**. When using the 0.6s time window, the accuracy is improved by 4% as a result of using the SE module. At other window lengths, the accuracy is also significantly improved. This demonstrates the effectiveness of the SE module. The SE module automatically adjusts the channel weights by learning the importance of different feature channels, thereby enhancing the attention to useful features and suppressing unimportant features. This feature relabeling can help the model to extract useful information from EEG signals more effectively, thereby improving the accuracy and robustness of the model. The addition of the SE module enables the network to adaptively learn the importance of features, and then improves the decoding performance of multi-target SSVEP signal decoding. We also investigated the effect of the position of the SE module. When the SE module is placed before the multi-scale convolution (referred to as Before SE), it merely recalibrates the extracted frequency and spatial information without integrating the temporal information of the EEG signals. Therefore, the SE module achieves effective feature reprocessing only when positioned after the multi-scale convolution.

C. Feature analysis

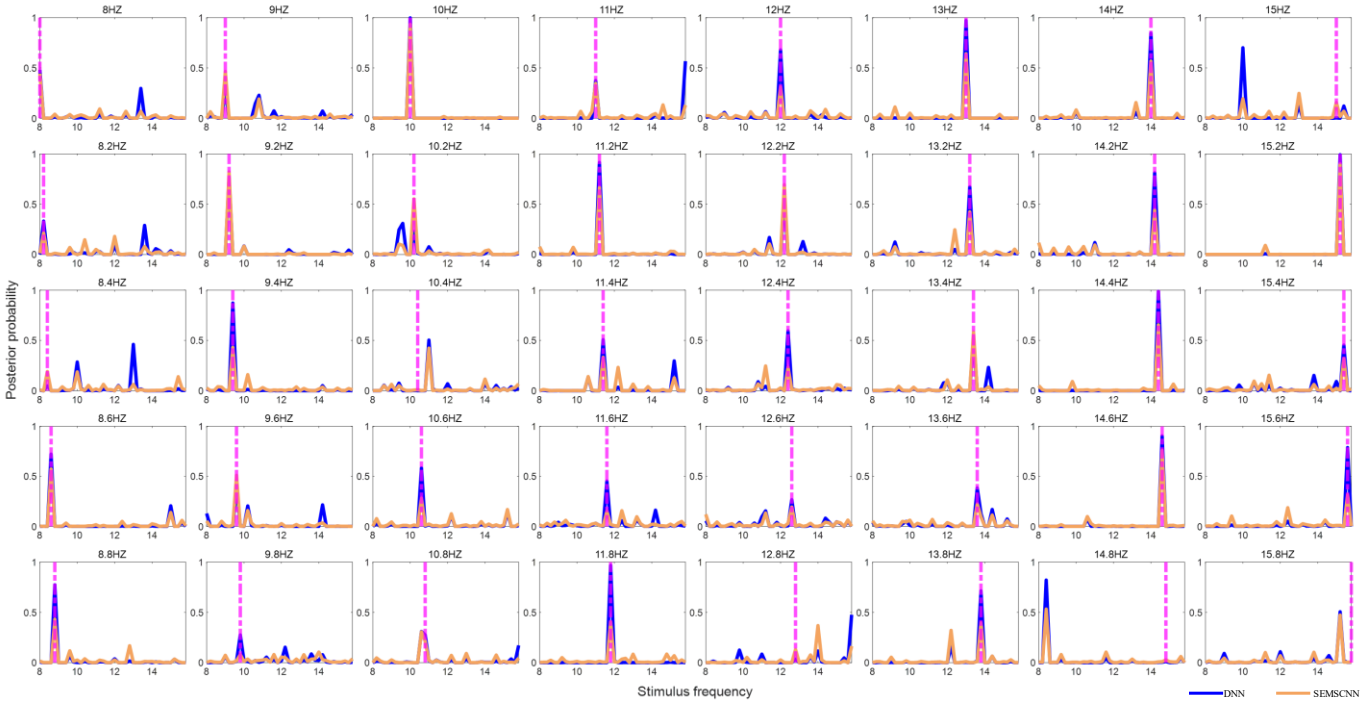
To further evaluate the effectiveness of our proposed method, feature analysis is performed. In the SEMSCNN network model, the posterior probability of an observation is obtained. This reflects how confident the model is that this observation belongs to a particular class. Specifically, the posterior probability is used to determine the final classification of the observations. The model will classify the observations into the class with the highest posterior probability. The posterior probability is used as an index to measure the importance of features. The datasets used in this study all contain 40 target categories, and the posterior probabilities of the DNN network and SEMSCNN network are compared on each target. Specific comparison results are shown in **Fig.6**. **Fig. 6(A)** presents the comparison results for one block of a participant on the Benchmark dataset. In **Fig.6(B)**, the comparison results on the BETA dataset are shown. For the target frequency from the figure, the probability of the proposed model is higher than that of the DNN model. This also shows that our proposed method . The size of the posterior probability can also indirectly reflect the importance of the features proposed by the model. The larger the posterior probability is, the more helpful the features proposed by the model are for classification, and the better the classification accuracy of the model.

D. Complexity analysis

To evaluate the computational complexity of the model presented in this paper, we further compared the model parameters, training time, and testing time of all network models, as shown in **TABLE V**. Here, we define the training time as the time required to train the global model and the testing time as the time needed to test the samples. During our evaluation, the data length for both datasets is set to 0.8 s. From **TABLE V**, we can observe that our proposed MSCNN and SEMSCNN models have more parameters than the baseline models. In terms of training time, SEMSCNN takes the longest, while DNN requires less time than MSCNN and SEMSCNN. For the testing time, SEMSCNN is the most time-consuming. MSCNN takes the least time for both training and testing. All models can perform effectively during the testing phase (under 60 ms), allowing for quick control of external IoT device



(A)



(B)

Fig.6. The feature analysis corresponding to all stimuli obtained by DNN and our proposed method from a representative participant. For the Benchmark dataset with 40 stimuli (A), and the Beta dataset with 40 stimuli (B), the data length is set as 0.2.

Energy consumption is determined by the time required to train the entire model and the power consumed by the computer. In **TABLE V**, we report the energy consumption during model training. SEMSCNN requires the most energy, while DNN requires less energy than MSCNN and SEMSCNN. For the testing phase, SEMSCNN uses the most energy.

V. Conclusion

In this study, we proposed a novel SEMSCNN model, and introduced the use of a SE module to learn the dependencies between features and to further improve the quality of feature extraction to enhance the recognition performance of SSVEPs for multi-stimulus frequency targets. We evaluated our proposed SEMSCNN model on two public datasets, Benchmark and Beta. Our experimental results show that the performance of our proposed method is significantly better than CORRCA, m-Extended-CCA, TSCORRCA, ms-eTRCA, and DNN. In addition to this, we also investigated the impact of different numbers of channels on the performance of our model. Our results show that 9 channels are optimal under the comprehensive consideration of accuracy and practical applications. Meanwhile, our ablation studies also indicated the suitability of our SEMSCNN model. By introducing the SE module, the extracted features are enhanced and the multi-target classification performance of SSVEP is further improved. This provides technical support for online SSVEP-based BCI applications. Furthermore, the model we proposed in this paper relies on specific participants, meaning the performance evaluation is conducted within-participants. Given the inherent variability among different participants, further evaluation is needed to assess cross-participant performance differences. In the future, we will attempt to further improve the model to make it more suitable for transfer learning ability and cross-actor classification scenarios, and then realize the control of IoT devices by different users using BCI technology.

References

[1] Lebedev, M. A., & Nicolelis, M. A. Brain–machine interfaces: past, present and future.*TRENDS in Neurosciences*, 29(9), 536-546, 2006.

[2] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces for communication and rehabilitation," *Nature Reviews Neurology*, vol. 12, no. 9, pp. 513-525, 2016.

[3] X. Gao, Y. Wang, X. Chen, and S. Gao, "Interface, interaction, and intelligence in generalized brain–computer interfaces," *Trends in cognitive sciences*, vol. 25, no. 8, pp. 671-684, 2021.

[4] Hu, H., Wang, Z., Zhao, X., Li, R., Li, A., Si, Y., ... & Xu, T. A Survey on Brain-Computer Interface-Inspired Communications: Opportunities and Challenges.*IEEE Communications Surveys & Tutorials*, doi: 10.1109/COMST.2024.3396847,2024.

[5] Zhang X, Yao L, Zhang S, et al. Internet of Things meets brain–computer interface: A unified deep learning framework for enabling human-thing cognitive interactivity[J]. *IEEE Internet of Things Journal*, 6(2): 2084-2092,2018.

[6] Fernández-Rodríguez, Álvaro, et al. "Effect of stimulus size in a visual ERP-based BCI under RSVP." *Sensors* 22.23: 9505,2022.

[7] J. Jin, R. Xu, I. Daly, X. Zhao, X. Wang, and A. J. I. T. o. C. Cichocki, MOCNN: A Multi-scale Deep Convolutional Neural Network for ERP-based Brain-Computer Interfaces,*IEEE Transactions on Cybernetics*,2024,DOI:10.1109/TCYB.2024.3390805

[8] Zhang, Y., Zhang, H., Gao, X., Zhang, S., & Yang, C.. Uav target detection for iot via enhancing erp component by brain–computer interface system. *IEEE Internet of Things Journal*, 10(19), 17243-17253. 2023.

[9] Zhang, Y., Zhang, H., Gao, X., Zhang, S., & Yang, C. (2023). Uav target detection for iot via enhancing erp component by brain–computer interface system.*IEEE Internet of Things Journal*,*10*(19), 17243-17253., 2023.

[10] Z. Wang, H. Hu, T. Zhou, T. Xu, and X. J. I. T. o. B. E. Zhao, "Average Time Consumption Per Character-a Practical Performance Metric for Generic Synchronous BCI Spellers," *IEEE Transactions on Biomedical Engineering*. doi: 10.1109/TBME.2024.3387469,2024.

[11] J. Jin *et al.*, Robust Similarity Measurement Based on a Novel Time Filter for SSVEPs Detection, *IEEE Transactions on Neural Networks and Learning Systems*, Aug;34(8):4096-4105,DOI:10.1109/TNNLS.2021.3118468,2023.

[12] Xiong, B., Wan, B., Huang, J., Li, F., Li, X., & Yang, P. Cross-Stimulus Transfer Method Using Common Impulse Response for Fast Calibration of SSVEP-Based BCIs. *IEEE Transactions on Instrumentation and Measurement*, (2024).

[13] Du, Y., Liu, J., Wang, X., & Wang, P. SSVEP-based emotion recognition for IoT via multiobjective neural architecture search. *IEEE Internet of Things Journal*, *9*(21), 21432-21443, (2022).

[14] Li, R., Zhao, X., Wang, Z., Xu, G., Hu, H., Zhou, T., & Xu, T. A Novel Hybrid Brain–Computer Interface Combining the Illusion-Induced VEP and SSVEP. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 4760-4772, (2023).

[15] Makary, M. M., Bu-Omer, H. M., Soliman, R. S., Park, K., & Kadah, Y. M. Spectral subtraction denoising preprocessing block to improve slow cortical potential based brain–computer interface. *Journal of Medical and Biological Engineering*,*38*, 87-98,(2018).

[16] Fumanal-Idocin, J., Wang, Y. K., Lin, C. T., Fernández, J., Sanz, J. A., & Bustince, H. Motor-imagery-based brain–computer interface using signal derivation and aggregation functions.*IEEE Transactions on Cybernetics*, *52*(8), 7944-7955, (2021).

[17] Zheng, Q., Wang, Y., & Heng, P. A. Multitask feature learning meets robust tensor decomposition for EEG classification. *IEEE Transactions on Cybernetics*, *51*(4), 2242-2252, 2019.

[18] Cho, J. H., Jeong, J. H., & Lee, S. W.. NeuroGrasp: Real-time EEG classification of high-level motor imagery tasks using a dual-stage deep learning framework. *IEEE Transactions on Cybernetics*, *52*(12), 13279-13292, 2021.

[19] Qi, F., Wu, W., Yu, Z. L., Gu, Z., Wen, Z., Yu, T., & Li, Y. Spatiotemporal-filtering-based channel

selection for single-trial EEG classification. *IEEE Transactions on Cybernetics*,*51*(2), 558-567, 2020.

[20] Huang, J., Yang, P., Xiong, B., Wang, Q., Wan, B., Ruan, Z., ... & Zhang, Z. Q. Incorporating neighboring stimuli data for enhanced SSVEP-based BCIs. *IEEE Transactions on Instrumentation and Measurement*,*71*, 1-9,2022.

[21] Li, Y., Xiang, J., & Kesavadas, T. Convolutional correlation analysis for enhancing the performance of SSVEP-based brain-computer interface.*IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(12), 2681-2690, 2020.

[22] Huang, J., Yang, P., Xiong, B., Wan, B., Su, K., & Zhang, Z. Q. Latency aligning task-related component analysis using wave propagation for enhancing SSVEP-based BCIs. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *30*, 851-859,2022.

[23] Zhang, Y., Li, Z., Xie, S. Q., Wang, H., Yu, Z., & Zhang, Z. Q. Multi-objective optimization-based high-pass spatial filtering for SSVEP-based brain–computer interfaces. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-9, 2022.

[24] Nakanishi, M., Wang, Y., Chen, X., Wang, Y. T., Gao, X., & Jung, T. P.. Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering*, *65*(1), 104-112, 2017

[25] Z. Lin, C. Zhang, W. Wu, and X. Gao, "Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs," *IEEE transactions on biomedical engineering*, vol. 53, no. 12, pp. 2610-2614, 2006.

[26] Q. Wei, S. Zhu, Y. Wang, X. Gao, H. Guo, and X. Wu, "A training data-driven canonical correlation analysis algorithm for designing spatial filters to enhance performance of SSVEP-based BCIs," *International journal of neural systems*, vol. 30, no. 05, p. 2050020, 2020.

[27] Lin, Z., Zhang, C., Wu, W., & Gao, X. Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs. *IEEE transactions on biomedical engineering*,*53*(12), 2610-2614, 2006.

[28] Bin, G., Gao, X., Wang, Y., Li, Y., Hong, B., & Gao, S. A high-speed BCI based on code modulation VEP. *Journal of neural engineering*, *8*(2), 025015,2011.

[29] Wei, Q., Zhu, S., Wang, Y., Gao, X., Guo, H., & Wu, X. A training data-driven canonical correlation analysis algorithm for designing spatial filters to enhance performance of SSVEP-based BCIs. *International journal of neural systems*, *30*(05), 2050020, 2020.

[30] Y. Zhang *et al.*, "Two-stage frequency recognition method based on correlated component analysis for SSVEP-based BCI," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 7, pp. 1314-1323, 2018.

[31] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung, and S. Gao, "High-speed spelling with a noninvasive brain–computer interface," *Proceedings of the national academy of sciences*, vol. 112, no. 44, pp. E6058-E6067, 2015.

[32] Y. Zhang *et al.*, "Hierarchical feature fusion framework for frequency recognition in SSVEP-based BCIs," *Neural Networks*, vol. 119, pp. 1-9, 2019.

[33] V. Bevilacqua *et al.*, "A novel BCI-SSVEP based approach for control of walking in virtual environment using a convolutional neural network," in *2014 international joint conference on neural networks (IJCNN)*, pp. 4121-4128,014.

[34] H. Cecotti, "A time–frequency convolutional neural network for the offline classification of steady-state visual evoked potential responses," *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1145-1153, 2011.

[35] H. Cecotti and A. Graeser, "Convolutional neural network with embedded Fourier transform for EEG classification," in *2008 19th International Conference on Pattern Recognition*, pp. 1-4, 2008.

[36] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*, pp. 1-6, 2017.

[37] Wu, J. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*,*5*(23), 495, 2017.

[38] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*,*9*, 611-629. 2018.

[39] Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS journal of photogrammetry and remote sensing*, *173*, 24-49. 2021.

[40] Cecotti, H., & Graser, A. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE transactions on pattern analysis and machine intelligence*, *33*(3), 433-445. 2010.

[41] S.-E. Moon, S. Jang, and J.-S. Lee, "Convolutional neural network approach for EEG-based emotion recognition using brain connectivity and its spatial information," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2556-2560, 2018.

[42] Zhang, Y., Zhang, X., Sun, H., Fan, Z., & Zhong, X. Portable brain-computer interface based on novel convolutional neural network. *Computers in biology and medicine*, *107*, 248-256. 2019.

[43] Kwak, N. S., Müller, K. R., & Lee, S. W. A convolutional neural network for steady state visual evoked potential classification under ambulatory environment. *PloS one*, *12*(2), e0172578, 2017.

[44] Zhang, X., Xu, G., Mou, X., Ravi, A., Li, M., Wang, Y., & Jiang, N. (2019). A convolutional neural network for the detection of asynchronous steady state

motion visual evoked potential. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(6), 1303-1311, 2019.

[45] Guney, O. B., Oblokulov, M., & Ozkan, H. A deep neural network for ssvep-based brain-computer interfaces. *IEEE transactions on biomedical engineering*,*69*(2), 932-944. 2021.

[46] Wang, Y., Chen, X., Gao, X., & Gao, S. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(10), 1746-1752.,2016.

[47] Liu, B., Huang, X., Wang, Y., Chen, X., & Gao, X. (2020). BETA: A large benchmark database toward SSVEP-BCI application. *Frontiers in neuroscience*, *14*, 627, 2020.

[48] Chen, X., Wang, Y., Nakanishi, M., Gao, X., Jung, T. P., & Gao, S. (2015). High-speed spelling with a noninvasive brain–computer interface. *Proceedings of the national academy of sciences*, *112*(44), E6058-E6067, 2015.

[49] Lin, F. C., Zao, J. K., Tu, K. C., Wang, Y., Huang, Y. P., Chuang, C. W., ... & Jung, T. P. (August). SNR analysis of high-frequency steady-state visual evoked potentials from the foveal and extrafoveal regions of human retina. In *2012 Annual international conference of the IEEE engineering in medicine and biology societ*, (pp. 1810-1814), 2012.

[50] Chen, X., Wang, Y., Gao, S., Jung, T. P., & Gao, X. Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface. *Journal of neural engineering*, *12*(4), 046008, 2015.

[51] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258, 2017.

[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141, 2018.

[53] D. P. Kingma and J. J. a. p. a. Ba, "Adam: A method for stochastic optimization," 2014.

[54] Nakanishi, M., Wang, Y., Wang, Y. T., & Jung, T. P. A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials. *PloS one*, *10*(10), e0140703, 2015.

[55] Wong, C. M., Wan, F., Wang, B., Wang, Z., Nan, W., Lao, K. F., ... & Rosa, A. Learning across multi-stimulus enhances target recognition methods in SSVEP-based BCIs. *Journal of neural engineering*, *17*(1), 016026. 2020.