Compact Artificial Neural Network Based on Task Attention for Individual SSVEP Recognition With Less Calibration

Ze Wang, *Member, IEEE*, Chi Man Wong, Boyu Wang, *Member, IEEE*, Zhao Feng, Fengyu Cong, *Senior Member, IEEE*, and Feng Wan, *Senior Member, IEEE*

***Abstract*— Objective: Recently, artificial neural networks (ANNs) have been proven effective and promising for the steady-state visual evoked potential (SSVEP) target recognition. Nevertheless, they usually have lots of trainable parameters and thus require a significant amount of calibration data, which becomes a major obstacle due to the costly EEG collection procedures. This paper aims to design a compact network that can avoid the over-fitting of the ANNs in the individual SSVEP recognition. Method: This study integrates the prior knowledge of SSVEP recognition tasks into the attention neural network design. First, benefiting from the high model interpretability of the attention mechanism, the attention layer is applied to convert the operations in conventional spatial filtering algorithms to the ANN structure, which reduces network connections between layers. Then, the SSVEP signal models and the common weights shared across stimuli are adopted to design constraints, which further condenses the trainable parameters. Results: A simulation study on two widely-used datasets demonstrates the proposed compact ANN structure with proposed constraints effectively eliminates redundant parameters. Compared to existing prominent deep neural network (DNN)-based and correlation analysis (CA)-based recognition algorithms, the proposed method reduces the trainable parameters by more than 90% and 80% respectively, and boosts the individual recognition performance by at least 57% and 7% respectively. Conclusion: Incorporating the prior knowledge of task into the ANN can make it more effective and efficient. The proposed ANN has a compact structure with less trainable parameters and thus requires less calibration with the prominent individual SSVEP recognition performance.**

***Index Terms*— Artificial neural network, attention layer, brain-computer interface, steady-state visual evoked potential.**

I. Introduction

BRAIN-COMPUTER interfaces (BCIs) provide the direct communication channels between human intentions and computer commands, and have found numerous applications and received increasing attention [1], [2], [3]. Due to high signal-to-noise ratios (SNRs) and high information transfer rates (ITRs), the steady-state visual evoked potential (SSVEP)-based BCI becomes the most prominent BCI paradigm [4], [5], [6].

Recognition method plays the important role of the performance in the SSVEP-based BCIs. Currently, the correlation analysis (CA)-based methods are widely applied for the SSVEP recognition and have achieved superior performance [7], [8], [9], [10]. These CA-based methods measure the similarities between the SSVEP reference or template signals and the weighted summation of multi-channel EEG signals, and classify the stimulus with the largest similarity as the target, where the weights of EEG channels are known as spatial filter [4]. One major difference between various CA-based methods is the type of reference or template signals: 1) The sine-cosine signals are widely used as the SSVEP reference signals, which was introduced in the standard canonical correlation analysis (sCCA) method [7], [8], [11], [12], [13], [14]; 2) The well-known template signals are the average calibration signals, which eliminates the interference from spontaneous EEG activities [8], [9], [10]. Another major difference between CA-based methods is the concatenating ways of the processed signals as well as the template or reference signals for determining spatial filters [4]: 1) In the sCCA, the individual template based CCA (itCCA) and the extended CCA (eCCA), the stimulus-specific

Manuscript received 26 September 2022; revised 24 April 2023; accepted 9 May 2023. Date of publication 29 May 2023; date of current version 2 June 2023. This work was supported in part by the Science and Technology Development Fund, Macau SAR under Grant 0045/2019/AFJ, Grant 0018/2019/AKP, and Grant 005/2022/ALC; in part by the University of Macau Research Committee under Grant MYRG 2017-00207-FST and Grant 2022-00197-FST; in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2023A1515010844; in part by the Self-Supporting Program of Guangzhou Laboratory under Grant SRPG22-007; and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grants Program. *(Corresponding author: Feng Wan.)*

Ze Wang is with the Macao Centre for Mathematical Sciences and the Respiratory Disease AI Laboratory on Epidemic Intelligence and Medical Big Data Instrument Applications, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China, and also with the Department of Electrical and Computer Engineering, Faculty of Science and Technology, University of Macau, Macau, China.

Chi Man Wong, Zhao Feng, and Feng Wan are with the Department of Electrical and Computer Engineering, Faculty of Science and Technology, University of Macau, Macau, China, and also with the Centre for Cognitive and Brain Sciences and the Centre for Artificial Intelligence and Robotics, Institute of Collaborative Innovation, University of Macau, Macau, China (e-mail: fwan@um.edu.mo).

Boyu Wang is with the Department of Computer Science, and the Brain Mind Institute, Western University, London, ON N6A 3K7, Canada.

Fengyu Cong is with the School of Biomedical Engineering, Faculty of Medicine, Dalian University of Technology, Dalian 116024, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2023.3276745, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2023.3276745

spatial filters are computed independently for different stimuli to maximize the correlation between the recorded EEG signals and the reference or template signals [7], [8], [15], [16]; 2) The multi-stimulus eCCA (ms-eCCA) concatenates reference and template signals of neighbor stimuli to calculate common spatial filters of multiple stimuli, and achieves superior performance [10]; 3) The task-related component analysis (TRCA), the ensemble TRCA (eTRCA), and the corresponding multi-stimulus scheme, namely multi-stimulus eTRCA (ms-eTRCA), consider the variances of different trials to further increase the recognition performance [9], [10]; 4) The task discriminant component analysis (TDCA) uses the discriminant analysis to consider the variances of different trials, includes the commonality of spatial filters to reduce the redundancy of spatial filters in (e)TRCA, and augments EEG data by including delayed signals as extra channels to expand the spatial filter to the spatio-temporal filter, which leads to the state-of-art performance [17]. Although these CA-based methods can provide the prominent performance, they compute the spatial weights without considering the recognition results, easily leading to non-optimized spatial filters. In addition, they only focus on exploring the spatial weights. The weights of other domains including the temporal and spectral domains are pre-defined, or trained individually with spatial filters, which may produce the error accumulation and thus reduce the performance.

Recently, several ANN based SSVEP recognition algorithms have been proposed. These algorithms provide the new perspectives of investigating SSVEP characteristics to overcome the issues of conventional CA-based algorithms. They achieve the end-to-end optimization of all model parameters simultaneously, which provides the global optimization and avoid the error accumulation [6], [18], [19], [20], [21], [22], [23]. The early studies are preliminary and limited by small number of targets. Several convolution neural network (CNN) and recurrent neural network (RNN) architectures were proposed to classify four or five stimulus targets, and can outperform the sCCA method and the conventional classifiers [18], [19], [21]. Recently, several DNN models were proposed for recognizing the large number of targets. As the number of targets increases, the calibration data sizes required by these DNN models also become prohibitively large. Guney et al. proposed one prominent DNN structure, which is termed as Guney-DNN in this study [6]. The Guney-DNN was firstly pre-trained over the calibration data from all subjects and then fine-turned for each subject. Another outstanding DNN structure, i.e., convolutional correlation analysis (Conv-CA), introduced in [22] only requires individual calibration data but still needs very long (full 5s) calibration data of all targets in all training blocks. The Guney-DNN and the Conv-CA both have been verified in the 40-stimulus-target SSVEP classification and can provide the state-of-art performance for the SSVEP recognition. Nevertheless, these DNN structures do not consider prior knowledge in conventional successful CA-based methods and SSVEP signal models, and thus contain lots of trainable parameters. Therefore, they require a large amounts of calibration signals from multiple subjects or with long signal lengths to avoid the over-fitting and the corresponding performance deterioration [24], [25]. In practical systems, because the calibration process is laborious and costly, such large and cross-domain calibration data is hard to be collected [10], [14], [26], [27], [28], [29]. Consequently, improving the practicality and the user experience of the SSVEP-based BCI systems poses a great challenge on distinguishing a large number of targets with small size, short and subject-specific calibration data.

This study explores the feasibility of adopting the prior knowledge of SSVEP tasks, including the SSVEP reference/template signals, the commonality of spatial filters, and the computations of spatial filters successfully used in these CA-based methods, to reduce the trainable parameters in the ANNs, and thus avoid the over-fitting issue. Moreover, the attention mechanism is adopted to design the network structure, which incorporates operations in conventional spatial filtering algorithms with the ANN architecture to reduce the connection numbers between layers. Inspired by the human attention, the attention layer makes the corresponding models concentrate on the most important information, which is similar as the idea behind the spatial filtering algorithms [30]. The spatial filtering algorithms and the attention layer both adopt weights to define the importance. The spatial filtering algorithms focus on exploring the importance of spatial information, and the attention layer explores the importance of information in multiple domains simultaneously. The attention-based ANNs have been adopted in various tasks, and demonstrated the superior performance, such as the image caption [31], the machine translation [32], the object detection [33], the action and facial expression recognition [34], [35]. The channel-wise and sample-wise attentions were widely adopted to explore the spatial dependence and the temporal dependence for the EEG signal decoding [30], [36], [37], [38], [39]. However, to our best knowledge, the relationships between the attention architecture and the conventional spatial filtering algorithms have not been explored in existing studies and have not been applied for the individual SSVEP recognition.

Compared to the conventional spatial filtering algorithms, the proposed ANN architecture is an end-to-end recognition model. It explores the importance of information in multiple domains including the temporal, spatial, feature, and spectral domains, and optimizes all trainable parameters simultaneously. Compared to the existing ANNs for the SSVEP recognition, the proposed ANN architecture incorporates prior knowledge of SSVEP tasks, including the SSVEP reference/template signals, the commonality of spatial filters across stimuli, and the operations in conventional spatial filtering algorithms, to condense trainable parameter dimensions and thus reduce the required calibration data size, making it suitable for the individual SSVEP recognition. Simulations conducted on two widely-used datasets show that the proposed ANN architecture outperforms the prominent CA-based and DNN-based methods, i.e., eCCA, TRCA, eTRCA, Guney-DNN, and Conv-CA, under the limited size of individual calibration data.

In the following Section II and Section III, we provide the preliminaries and introduce the proposed task attention based individual SSVEP recognition method. After we present the performance evaluation in Section IV and the discussions in Section V, the conclusion is given in Section VI.

II. Preliminaries

A. Unified Framework of Spatial Filtering Algorithms and Ensemble Technique of eCCA

By leveraging the frequency- and phase-locking characteristics of the SSVEP signals, the SSVEP-based BCIs encode commands by visual flickers with different flashing frequencies and phases, and decode users' intentions by identifying dominant frequency components [5]. In the conventional SSVEP recognition methods, the core process is to determine the suitable spatial filters that combine EEG signals in multiple channels to improve the signal quality. The obtained spatial filters can enhance the SNRs and protects the extraction of reliable features. Wong et. al summarized various prominent spatial filtering algorithms [4], and unified them as a generalized eigenvalue problem (GEP) that is

$$\mathbf{D}\mathbf{T}^T\mathbf{T}\mathbf{D}^T\mathbf{W}^T = \mathbf{B}\mathbf{W}^T\Delta, \tag{1}$$

where $\mathbf{D}$ denotes processed EEG signals, $\mathbf{W}$ denotes the eigenvectors and works as the spatial filters, Δ denotes the corresponding eigenvalues, $\mathbf{B}$ is the covariance matrix of $\mathbf{D}$, and $\mathbf{T}$ denotes the temporal filter that is pre-defined and computed by the SSVEP reference/template signals.

In addition to the spatial filter approach, the ensemble technique used in the eCCA also can significantly improve the recognition performance, and widely used in the SSVEP recognition algorithms [8], [10], [16], [29]. The eCCA combines the ideas of the sCCA and the itCCA. It finds 3 types of spatial filters by solving (1) with SSVEP reference and template signals respectively: 1) $\mathbf{D} = \mathbf{X}$ and $\mathbf{T} = \mathbf{Q}_{\mathbf{R}_i}$, 2) $\mathbf{D} = \mathbf{X}$ and $\mathbf{T} = \mathbf{Q}_{\overline{\mathbf{X}}_i}$, as well as 3) $\mathbf{D} = \overline{\mathbf{X}}_i$ and $\mathbf{T} = \mathbf{Q}_{\mathbf{R}_i}$, where $\mathbf{X}$ denotes the EEG signals that need to be recognized, $\mathbf{R}_i$ and $\overline{\mathbf{X}}_i$ denote the SSVEP reference and template signals respectively, as well as $\mathbf{Q}_{\mathbf{R}_i}$ and $\mathbf{Q}_{\overline{\mathbf{X}}_i}$ denote the orthogonal matrices obtained from the QR factorization of $\mathbf{R}_i$ and $\overline{\mathbf{X}}_i$ respectively. Then, the eCCA ensembles all results obtained by these spatial filters [16].

B. SSVEP Reference and Template Signals

According to the hypothesis that the SSVEP signals are the output of a linear system with the corresponding stimulus signals as the input, the well-known SSVEP reference signal is a set of sine-cosine signals, which can be presented as

$$\mathbf{R}_i = \begin{bmatrix} \sin(2\pi f_i t + \theta_i) \\ \cos(2\pi f_i t + \theta_i) \\ \vdots \\ \sin(2\pi N_h f_i t + N_h\theta_i) \\ \cos(2\pi N_h f_i t + N_h\theta_i) \end{bmatrix}, \tag{2}$$

where N_h denotes the total number of harmonic components.

To avoid the interference from the spontaneous EEG activities, the individual templates constructed by calibration data are also commonly applied for the SSVEP recognition. The widely used SSVEP template is the average signal of all calibration trials, which can be expressed as

$$\overline{\mathbf{X}}_i = \frac{1}{N_t N_b}\sum_{n_b=1}^{N_b}\sum_{n_t=1}^{N_t}\mathbf{X}_{i,n_b,n_t}, \tag{3}$$

where $\mathbf{X}_{i,n_b,n_t}$ denotes the calibration data of the i-th stimulus in the n_t-th calibration trial of the n_b-th training block, N_t denotes the number of calibration trials in one training block, and N_b denotes the total number of training blocks.

C. Scaled Dot-Product Attention

An attention function can be described as the weighted summation of correlations of the queries and the keys, where the weights are the corresponding values [40]. In practice, the scaled dot-product attention architecture was proposed in [32]. Specifically, the output of the attention layer can be computed as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_\mathbf{K}}}\right)\mathbf{V}, \tag{4}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ denote the packed sets of queries, keys, and values, and $d_\mathbf{K}$ is the dimension of keys. The structure of the scaled dot-product attention is illustrated in Fig. 1(a).

D. Statistical Analysis Method

To evaluate the significance of the performance differences between recognition methods, trainable parameter numbers, training block numbers, and channel numbers, the paired t-test is applied, while the p-values are corrected by the Bonferroni correction [6]. For the N paired t-tests containing total N comparisons, we denote the significance of an observed difference as "*" if the p-value is less than $0.05/N$, "**" if the p-value is less than $0.01/N$, "***" if the p-value is less than $0.001/N$, and "n.s." if the p-value is larger than $0.05/N$.

III. Materials and Methods

A. SSVEP Datasets

The proposed ANN architecture is validated on two public datasets:

1) The benchmark dataset was collected from 35 healthy subjects participating in the SSVEP-based BCI experiment [41]. This experiment uses a 40-target BCI speller and a sampled sinusoidal stimulation method to present visual stimuli where the luminance of the screen is controlled by the stimulus sequence sampled from $0.5 + 0.5\sin(2\pi f_i t + \theta_i)$ where f_i and θ_i denote the stimulus frequency and phase of the i-th stimulus, respectively. The stimulus frequencies start from 8Hz to 15.8Hz with 0.2Hz interval. The stimulus phases range from 0 to 1.5π with 0.5π interval. All stimuli have the same size and are evenly distributed. This experiment includes 6 blocks, with each block consisting of 40 trials corresponding to 40 stimuli. Every trial starts with a 0.5s target cue. Then, all stimuli are flickered on the screen for 5s. Finally, the screen goes blank for 0.5s break before next trial.

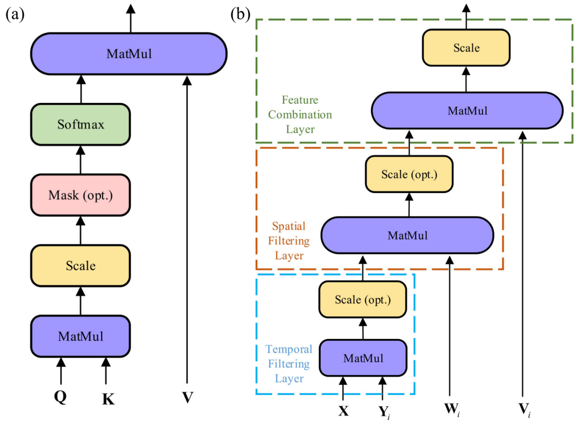

Fig. 1. Attention architectures. (a) Scaled dot-product attention. (b) Proposed task attention.

2) The BETA dataset was recorded from 70 healthy subjects [42]. The stimulus design is similar as that in the benchmark dataset, but the stimuli are distributed as the traditional QWERT keyboard. Each experiment consists of 4 blocks. The stimulation lasts 2s for the first 15 subjects and 3s for the remaining subjects. Because the experiments are conducted outside of the laboratory, the signal qualities are lower than that in the benchmark dataset, leading to more challenges of the target recognition in the BETA dataset.

EEG signals of these two datasets were recorded from 64 channels and down-sampled to 250Hz. A notch filter at 50Hz was applied to remove the power-line noise. EEG data in each trial were filtered using a band-pass filter with the lower and higher cut-off frequencies of 7Hz and 90Hz, respectively. The average visual latencies are approximately estimated as 0.14s in the benchmark dataset and 0.13s in the BETA dataset [41], [42].

B. Task Attention Architecture

The proposed task attention architecture illustrated in Fig. 1(b) integrates (1) and (4) together. It contains three layers for the i-th stimulus:

1) The first layer acts as the temporal filtering in the conventional SSVEP recognition methods, which is called temporal filtering layer in this study. The temporal filtering layer measures the similarities of the processed EEG signals $\mathbf{X}$ and the kernels $\mathbf{Y}_i$, which is similar as the multiplication of the queries $\mathbf{Q}$ and the keys $\mathbf{K}$ in the scaled dot-product attention shown in Fig. 1(a). The queries and the keys in a conventional self-attention architecture are projected from the input, where the projection matrices are learned from the training process. On the contrary, according to the frequency- and phase-locked characteristics of the SSVEP signals and the unified framework of the spatial filtering algorithms, the kernels of the proposed first layer are pre-defined as the SSVEP reference and/or template signals to reduce the number of trainable parameters.
2) The second layer computes the channel-wise weights $\mathbf{W}_i$ of the outputs from the first layer, which is called spatial filtering layer in this study. The channel-wise weights are similar as values $\mathbf{V}$ of the attention layer shown in Fig. 1(a). The values in the conventional self-attention are also projected from the input but the channel-wise weights in the spatial filtering layer are trained by the calibration signals.
3) After combining temporal features across channels in the second layer, the third layer explores the importance of these combined features inspired by the unified frame work of spatial filtering algorithms. The feature-wise weights $\mathbf{V}_i$ work similar as the channel-wise weights $\mathbf{W}_i$ but are performed on different dimensions, and are also optimized by the calibration data.

It should be noted that, due to the time-varying characteristics of EEG signals, the ranges of these features may be varied in different trials, and thus should be scaled to a unified range. Following the unified framework of the spatial filtering algorithms, such scaling operation is operated after the third layer in this study.

C. Multi-Head Task Attention Architecture

Following the ensemble technique of the eCCA introduced in Section II-A, this study proposes to use the multi-head task attention layer, instead of one single attention layer. The multi-head task attention layer is similar as the multi-head attention proposed in [32]. It incorporates multiple task attention layers with different kernels in the temporal filtering layer as well as different trainable weights in the spatial filtering layer and the feature combination layer, which jointly projects SSVEP signals into different representation subspaces and then properly combines the information in these subspaces. The whole architecture is shown in Fig. 2(a).

In the temporal filtering layer of the i-th stimulus, both SSVEP reference signals $\mathbf{R}_i \in \mathbb{R}^{H\times T}$ and SSVEP template signals $\overline{\mathbf{X}}_i \in \mathbb{R}^{C\times T}$ are applied as kernels $\mathbf{Y}_i$, where C, H, and T denote the EEG channel number, the harmonic number, and the sampling number. The projected temporal features of the input signal $\mathbf{X} \in \mathbb{R}^{C\times T}$ can be expressed as

$$\mathbf{F}_i^{(\mathrm{ref})} = \mathbf{X}\mathbf{R}_i^T \ \in \mathbb{R}^{C\times H} \text{ and } \mathbf{F}_i^{(\mathrm{temp})} = \mathbf{X}\overline{\mathbf{X}}_i^T \ \in \mathbb{R}^{C\times C}. \quad (5)$$

Four pairs of trainable weights in the spatial filtering layer and the feature combination layer explores the importance of the spatial and feature-based information from different aspects. The first pair of trainable weights $\mathbf{W}_{i,1}$ and $\mathbf{V}_{i,1}$ is based on two assumptions: 1) The spatial weights of $\mathbf{F}_i^{(\mathrm{ref})}$ and $\mathbf{F}_i^{(\mathrm{temp})}$ are same; 2) The spatial distributions of the recorded EEG signals and the template signals are same. This pair of trainable weights leads to two outputs from the feature combination layer, which can be represented as

$$\mathbf{O}_{i,1} = \frac{\mathbf{W}_{i,1}\mathbf{F}_i^{(\mathrm{ref})}\mathbf{V}_{i,1}^T}{\sqrt{D_{i,1}}} \text{ and } \mathbf{O}_{i,2} = \frac{\mathbf{W}_{i,1}\mathbf{F}_i^{(\mathrm{temp})}\mathbf{W}_{i,1}^T}{\sqrt{D_{i,2}}}. \quad (6)$$

Two assumptions for the second and third pairs of trainable weights $\mathbf{W}_{i,0}$ and $\mathbf{V}_{i,0}$ as well as $\mathbf{W}_{i,2}$ and $\mathbf{V}_{i,2}$ are opposite to the assumptions of the first pair of trainable weights: 1) The spatial weights of $\mathbf{F}_i^{(\mathrm{ref})}$ and $\mathbf{F}_i^{(\mathrm{temp})}$ are independent; 2) The spatial distributions of the recorded EEG signals and

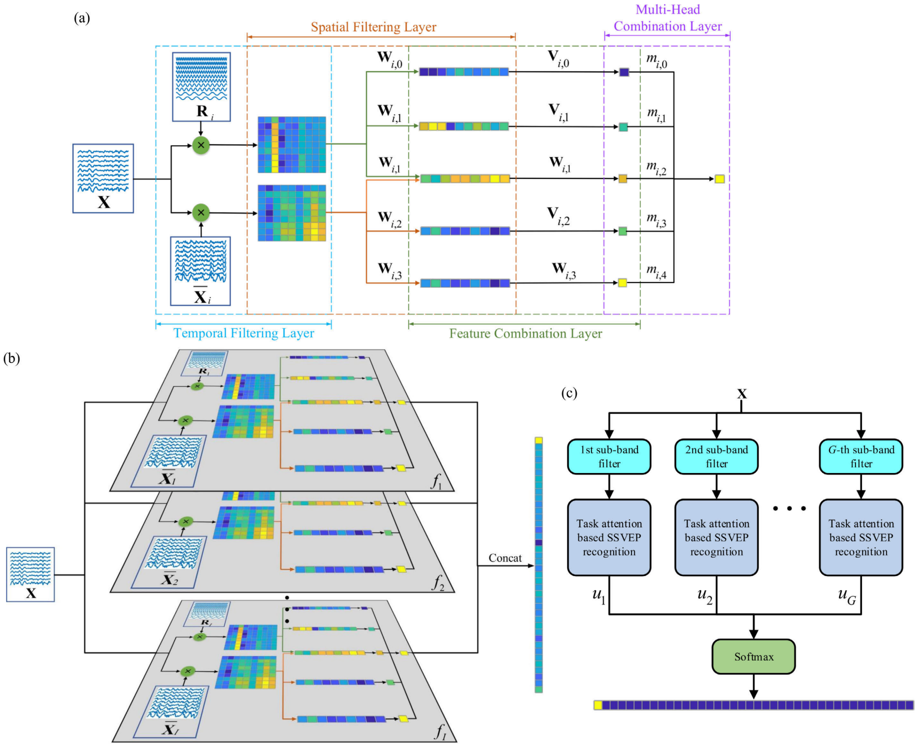

Fig. 2. Proposed ANN architecture. (a) Multi-head task attention. (b) Task attention based SSVEP recognition architecture. (c) Task attention based SSVEP recognition with filter-bank approach.

the template signals are also independent. Based on these assumptions, two outputs can be expressed as

$$\mathbf{O}_{i,0} = \frac{\mathbf{W}_{i,0}\mathbf{F}_i^{(\text{ref})}\mathbf{V}_{i,0}^T}{\sqrt{D_{i,0}}} \text{ and } \mathbf{O}_{i,3} = \frac{\mathbf{W}_{i,2}\mathbf{F}_i^{(\text{temp})}\mathbf{V}_{i,2}^T}{\sqrt{D_{i,3}}}. \tag{7}$$

The fourth pair of trainable weights $\mathbf{W}_{i,3}$ and $\mathbf{V}_{i,3}$ assume that the spatial distributions of the recorded EEG signals and the template signals are same and not related to the spatial weights of $\mathbf{F}_i^{(\text{ref})}$. The corresponding output can be computed as

$$\mathbf{O}_{i,4} = \frac{\mathbf{W}_{i,3}\mathbf{F}_i^{(\text{temp})}\mathbf{W}_{i,3}^T}{\sqrt{D_{i,4}}}. \tag{8}$$

Inspired by the unified framework of the spatial filtering algorithms, the scaling factors $\{D_{i,d}\}_{d=0,1,2,3,4}$ are calculated by the variances of the input EEG signal and the corresponding reference or template signals:

$$D_{i,0} = \mathbf{W}_{i,0}\mathbf{X}\mathbf{X}^T\mathbf{W}_{i,0}^T\mathbf{V}_{i,0}\mathbf{R}_i\mathbf{R}_i^T\mathbf{V}_{i,0}^T, \tag{9}$$

$$D_{i,1} = \mathbf{W}_{i,1}\mathbf{X}\mathbf{X}^T\mathbf{W}_{i,1}^T\mathbf{V}_{i,1}\mathbf{R}_i\mathbf{R}_i^T\mathbf{V}_{i,1}^T, \tag{10}$$

$$D_{i,2} = \mathbf{W}_{i,1}\mathbf{X}\mathbf{X}^T\mathbf{W}_{i,1}^T\mathbf{W}_{i,1}\overline{\mathbf{X}}_i\overline{\mathbf{X}}_i^T\mathbf{W}_{i,1}^T, \tag{11}$$

$$D_{i,3} = \mathbf{W}_{i,2}\mathbf{X}\mathbf{X}^T\mathbf{W}_{i,2}^T\mathbf{V}_{i,2}\overline{\mathbf{X}}_i\overline{\mathbf{X}}_i^T\mathbf{V}_{i,2}^T, \tag{12}$$

$$D_{i,4} = \mathbf{W}_{i,3}\mathbf{X}\mathbf{X}^T\mathbf{W}_{i,3}^T\mathbf{W}_{i,3}\overline{\mathbf{X}}_i\overline{\mathbf{X}}_i^T\mathbf{W}_{i,3}^T. \tag{13}$$

Then, a multi-head combination architecture is designed to explore the importance of different task attention layers, which combines outputs $\{\mathbf{O}_{i,d}\}_{d=0,\cdots,4}$ from multiple task attention architectures. Following the ensemble technique proposed in the eCCA [8], the multi-head combination layer can be expressed as

$$\mathbf{O}_i^{\text{multi}} = \sum_{d=0}^{4}\left[m_{i,d} \cdot \text{sign}\left\{\mathbf{O}_{i,d}\right\} \cdot \mathbf{O}_{i,d}^2\right], \tag{14}$$

where $m_{i,d}$ denotes the trainable weight of the d-th output of the task attention architecture for the i-th stimulus.

Compared to the fully connection layer, the proposed architecture reduces the connection numbers to condense the trainable parameter number. Furthermore, The trainable parameter dimensions in the spatial filtering layer, the feature combination layer, and the multi-head combination layer are further limited from two aspects. Firstly, inspired by the sCCA, the parameter dimensions for all $i \in \{1, 2, \cdots, I\}$ are set as

$$\left\{\mathbf{W}_{i,n_w}\right\}_{n_w=0,1,2,3} \text{ and } \mathbf{V}_{i,2} \in \mathbb{R}^{1\times C}, \text{ and}$$
$$\left\{\mathbf{V}_{i,n_w}\right\}_{n_w=0,1} \in \mathbb{R}^{1\times H}. \tag{15}$$

Secondly, the commonalities of the spatial filters and the ensemble weights across stimuli introduced in [8] and [10] are adopted, which can be expressed as

$$\begin{aligned} \mathbf{W}_{1,w} &= \mathbf{W}_{2,w} = \cdots = \mathbf{W}_{I,w} \;\forall\; w = 0, 1, 2, 3, \\ \mathbf{V}_{1,v} &= \mathbf{V}_{2,v} = \cdots = \mathbf{V}_{I,v} \;\forall\; v = 0, 1, 2, \\ m_{1,d} &= m_{2,d} = \cdots = m_{I,d} \;\forall\; d = 0, 1, 2, 3, 4. \end{aligned} \tag{16}$$

D. Multi-Head Task Attention Architecture Based SSVEP Recognition Algorithm

In this study, the multi-head task attention architectures corresponding to all stimuli are parallel concatenated together as shown in Fig. 2(b). Moreover, the filter-bank approach that has been successfully applied in existing prominent SSVEP recognition methods is also adopted in this study as shown in Fig. 2(c). The outputs from sub-bands are regarded as the spectral information and integrated by

$$\mathbf{O}_i^{\text{filterbank}} = \sum_{g=1}^{G} u_g \mathbf{O}_{i,g}^{\text{multi}}, \tag{17}$$

where $\mathbf{O}_{i,g}^{\text{multi}}$ denotes the output of the g-th sub-band for the i-th stimulus, u_g denotes the weights of the g-th sub-band, and G denotes the total number of sub-bands. The sub-band weights present the importance of the spectral information. In addition, following the eCCA, the weights in the multi-head combination layers of all sub-bands are limited as the same values to condense the trainable parameters.

Finally, the softmax layer unifies the classification results of all stimuli together. The i-th element of the output from the softmax layer is computed as $\mathbf{O}_i^{\text{softmax}} = e^{\mathbf{O}_i^{\text{filterbank}}} \Big/ \sum_{k=1}^{I} e^{\mathbf{O}_k^{\text{filterbank}}}$. The stimulus providing the largest output of the softmax layer is regarded as the recognition result.

E. Training Process

This study uses the gradient descent to minimize the categorical cross-entropy loss. The whole training process contains several iterative epochs. In each epoch, calibration signals and corresponding labels of all training blocks are integrated together as pairs of inputs and outputs in one batch to update trainable parameters. For one pair of input and output, supposing whole training blocks are applied to construct the SSVEP templates following (3), the model parameters will include the information from the input, which easily leads to the over-fitting issue. To avoid this problem, the computation of the SSVEP template in the training process for the input and output pair of the i-th stimulus in the n_b-th training block does not involve the calibration data of the n_b-th training block, which follows

$$\overline{\mathbf{X}}_{i,n_b} = \frac{1}{N_t\,(N_b - 1)} \left(\sum_{j=1}^{N_b} \sum_{m=1}^{N_t} \mathbf{X}_{i,j,m} - \sum_{m=1}^{N_t} \mathbf{X}_{i,n_b,m} \right). \tag{18}$$

For the recognition or testing process, since the new incoming blocks are not included in the training blocks, the SSVEP templates are still computed by (3). The related codes are available at https://github.com/pikipity/Compact-Artificial-Neural-Network-SSVEP.

IV. Results

A. Classification Performance

The classification performance is verified by the leave-one-block-out cross validation. In all blocks, one block is used for testing, and remaining blocks are adopted to construct kernels in the temporal filtering layer and train model parameters, which ensures that the training and testing datasets are from different blocks, and thus avoids using the information outside the training dataset to create the recognition model. The entire evaluation is repeated to test all blocks. The classification performance is evaluated by the classification accuracy and the ITR computed by ITR $= 60/T \cdot \left\{\log_2 I + P\log_2 P + (1-P)\log_2\left[(1-P)/(I-1)\right]\right\}$ where P is the classification accuracy, and T is the total time of each detection that is equal to the summation of the shifting visual attention time (0.5s in both datasets), the visual latency (0.14s in the benchmark dataset and 0.13s in the BETA dataset), the signal length for the target recognition, and the actual computational time of each detection. Following [41] and [42], the EEG signals from 9 selected channels (Pz, PO-z/3/4/5/6, O-z/1/2) are utilized in this section. In addition, following [11], the total sub-band number G is set to 5 in this study. The lower and upper cut-off frequencies of the g-th sub-band are set to $(8 \cdot g)$Hz and 90Hz. The classification performance of the proposed task attention based SSVEP recognition method is compared with that of four prominent CA-based individual SSVEP recognition methods, i.e., the sCCA, the eCCA, the TRCA and the eTRCA, and two prominent DNN-based SSVEP recognition methods, i.e., the Guney-DNN and the Conv-CA, which is illustrated in Fig. 3. For each signal length, six paired t-tests are conducted.

The proposed ANN architecture can deliver the best performance in comparison with all other algorithms in terms of the accuracy and the ITR in both datasets. It can provide 182.046 bits/min (85.488% accuracy) and 137.309 bits/min (74.982% accuracy) for the benchmark dataset and the BETA dataset, respectively. These maximum ITRs are achieved in 0.5s and 0.6s signal lengths for the benchmark dataset and the BETA dataset, respectively. In the benchmark dataset, when the calibration data sizes are small (short signal lengths, i.e., 0.25s and 0.5s), the proposed ANN architecture can significantly outperform other methods. In the BETA dataset, due to lower signal qualities and smaller training block numbers, the recognition performance of all methods is lower than that in the benchmark dataset. In this case, the proposed ANN architecture performs significantly better than other methods for all signal lengths. Such impressive results under the limited calibration data size and the unsatisfactory signal quality provides reassurance about the robustness of our proposed method. It should be noted that, since this study focuses on the individual SSVEP recognition, only individual calibration signals whose signal lengths are same as those of testing signals are utilized in the training processes for all methods to achieve fair comparisons. In other words, the calibration data size in this study is much smaller than that in [6] and [22]. Therefore, the recognition results of the Conv-CA and the Guney-DNN are much lower than the results presented in [6] and [22].

B. Trainable Parameter Size

The effect of the trainable parameter number is verified in this section. Table I compares the trainable parameter numbers and the maximum average ITRs of the proposed ANN architecture as well as the prominent CA-based and DNN-based methods. We can observe that the proposed ANN architecture needs the smallest trainable parameter number and can achieve the highest maximum ITR. Compared to the eTRCA that needs the smallest trainable parameter number

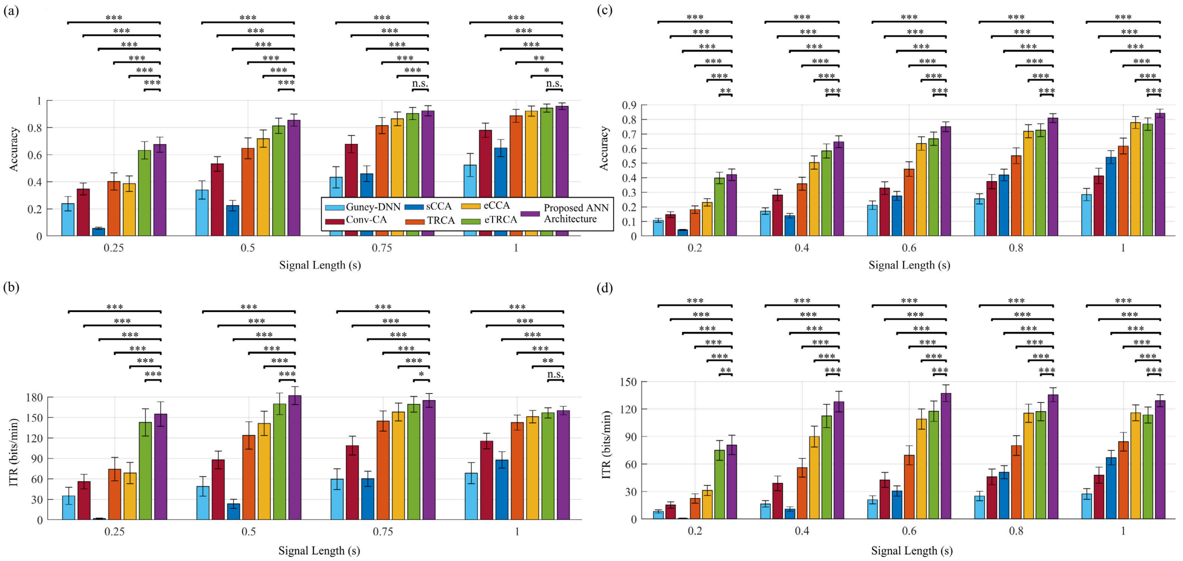

Fig. 3. Classification performance of the proposed ANN architecture, the prominent CA-based and DNN-based SSVEP recognition methods for different signal lengths. (a) Classification accuracy for the benchmark dataset. (b) ITR for the benchmark dataset. (c) Classification accuracy for the BETA dataset. (d) ITR for the BETA dataset.

TABLE I

COMPARISONS OF TRAINABLE PARAMETER NUMBERS AND MAXIMUM ITRS($C = 9$, $H = 5$, $G = 5$, AND $I = 40$)

Method	Trainable parameter number[a]	Maximum ITR in benchmark dataset (bits/min)[b]	Maximum ITR in BETA dataset (bits/min)[b]
sCCA	3800	87.659 (1.00s)	66.987 (1.00s)
eCCA	11200	158.070 (0.75s)	115.874 (1.00s)
TRCA	1800	144.773 (0.75s)	84.367 (1.00s)
eTRCA	1800	169.973 (0.50s)	117.635 (0.60s)
Guney-DNN	774286	68.519 (1.00s)	27.482 (1.00s)
Conv-CA	323086	115.421 (1.00s)	47.930 (1.00s)
Proposed ANN	**335**	**182.046** (0.50s)	**137.309** (0.60s)

[a] The detailed computation processes of the trainable parameters are shown in Section S.I of the supplementary material.

[b] The time values in brackets denote the corresponding signal lengths.

and provides the best performance in these four CA-based methods, the proposed ANN architecture reduces the trainable parameter number by 81.389%, and improves the maximum ITRs by 7.103% and 16.725% for the benchmark and BETA datasets, respectively. Compared to the Conv-CA that needs the smallest trainable parameter number and provides the best performance in these two DNN-based methods, the proposed ANN architecture reduces the trainable parameter number by 99.896%, and improves the maximum ITRs by 57.724% and 186.478% for the benchmark and BETA datasets, respectively.

The pre-defined kernels of the temporal filtering layers introduced in Section III-B and the constrains of other layers shown in (16) have large contributions on limiting dimensions of trainable parameters. Fig. 4 illustrates the comparisons of the proposed ANN architecture with and without limitations on parameter dimensions, including the pre-defined kernels in the temporal filtering layers and the constrains in other layers. When the parameters in all layers are unlimited, the trainable parameter numbers are 248630, 488030, 723630, and 963030 at the signal lengths of 0.25s, 0.5s, 0.75s, and 1s, respectively. When only parameters in the temporal filtering layer is limited, the trainable parameter numbers are 13030 at all signal lengths. When parameters in all layers are limited, the trainable parameter numbers are 335 at all signal lengths. The benchmark dataset is adopted due to good signal quality and enough number of training blocks. Two paired t-tests are conducted in each signal length. The proposed ANN architecture with the limitations on parameters of all layers can provide the best recognition performance. Fig. 5 shows the testing and training accuracies in all training epochs. To show the performance deterioration when removing limitations on parameter sizes and the lower limit of performance, the signal length is set as 0.25s where the recognition accuracy is low. Although the testing accuracy of the proposed ANN architecture both with and without the limitations on parameter dimensions can converge to very large values, the differences between the training and testing results become large as the sizes of trainable parameters increase. More results can be found in Section S.II of the supplementary material.

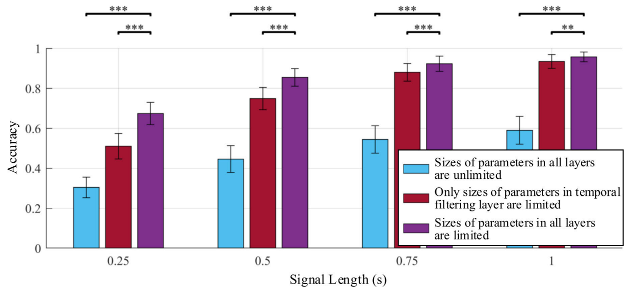

Fig. 4. Comparisons of the proposed ANN architecture with and without limitations on parameter dimensions.

C. Calibration Data Size

In the individual SSVEP recognition, the calibration data size normally is determined by the number of calibration blocks and channel numbers. Effects of these two parameters on the calibration performance are evaluated in this section. Because the benchmark dataset contains more blocks, it is

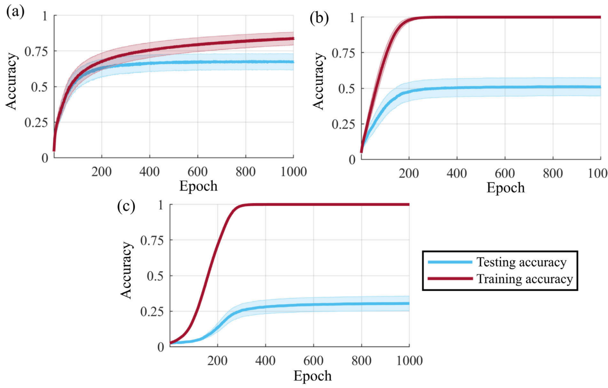

Fig. 5. Convergences of testing and training accuracies delivered by the proposed ANN architecture (0.25s signal length for the benchmark dataset). (a) Sizes of parameters in all layers are limited. (b) Only sizes of parameters in the temporal filtering layer is limited. (c) Sizes of parameters in all layers are unlimited.

TABLE II

Computational Time of the Proposed ANN Architecture and the Prominent CA-Based Methods for the Recognition Process of One Block Signals in the Benchmark Dataset($C = 9$, $H = 5$, $G = 5$, and $I = 40$)

Method	sCCA	eCCA	TRCA	eTRCA	Proposed ANN
Computational Time (s)	2.769 ± 0.098	7.871 ± 0.301	0.606 ± 0.029	0.753 ± 0.035	0.668 ± 0.062

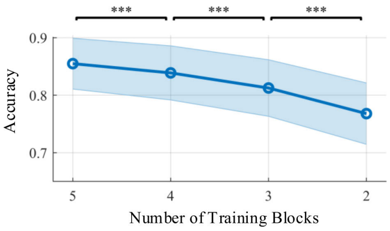

Fig. 6. Classification accuracy of the proposed ANN architecture for different numbers of training blocks (0.5s signal length for the benchmark dataset).

adopted to evaluate the effects of different training block numbers.

Fig. 6 illustrates the classification accuracy of the proposed ANN architecture for different training block numbers. To show the performance improvement as the number of training blocks increases and the upper limit of performance, the window length is set as 0.5s where the proposed ANN architecture can achieve the maximum ITR as shown in Fig. 3. Three paired t-tests are conducted to analyze the performance differences between different numbers of training blocks. All p-values are less than $0.001/3$, indicating that the differences of the recognition performance between different numbers of training blocks are significant. More results can be found in Section S.III of the supplementary material.

Fig. 7 illustrates the classification accuracy of the proposed ANN architecture and the prominent CA-based methods for different channel numbers. The recognition performance of the proposed ANN architecture and the prominent three CA-based methods, i.e., the eCCA, the TRCA, and the eTRCA, is evaluated for 9 (Pz, PO-z/3/4/5/6, O-z/1/2), 19 (P-z/1/2/3/4/5/6/7/8, PO-z/3/4/5/6/7/8, O-z/1/2), and 32 (all channels in occipital, parietal, and central-parietal regions) channels. To show the robustness of the proposed method and the lower limit of performance, the window length is set as 0.25s and 0.4s for the benchmark dataset and the BETA dataset respectively, where the recognition accuracies of all methods are low. For each method, two paired t-tests are conducted to analyze the performance differences between different EEG channel numbers. It can be seen that, as the EEG channel number increases, the recognition accuracies of the TRCA and the eCCA have the significant deduction, especially for the eCCA. The performance of the eTRCA and the proposed ANN could be significantly improved when the channel number increases. For each channel number, three paired t-tests are conducted. Each t-test evaluates the performance difference between the proposed ANN architecture and one of two CA-based methods. No matter how many EEG channel number, the proposed ANN architecture always can significantly outperform other three methods. More results can be found in Section S.IV of the supplementary material.

V. Discussions

A. Compact ANN and Less Calibration

One key contribution of this study is that the prior knowledge of the SSVEP task is integrated to the attention neural network, which significantly reduce the trainable parameters to avoid the over-fitting issue and leads to the impressive recognition performance as shown in Fig. 3 and Table I. More specifically, the reduction of the trainable parameters is mainly due to 2 designs: 1) According to the operations in conventional spatial filtering algorithms, the connection numbers between layers are limited; 2) Inspired by the SSVEP reference/template signals and the commonality of spatial filters across stimuli, the constraints are designed to condense the dimensions of trainable parameters. Removing any limitations in these two designs exacerbates the over-fitting issue, which reduces the generalization ability of the proposed model, and thus leads to large gaps between the training and testing accuracy as shown in Fig. 5. The over-fitting issue is the key reason of the performance deterioration when the number of trainable parameters increases illustrated in Fig. 4. These results suggest that condensing the ANN model size can effectively avoid the over-fitting and improve the recognition performance under the limited calibration data in the individual SSVEP recognition.

Although the proposed method requires much less calibration data than other ANN methods, the calibration data size still has the large effect on the recognition performance. Fig. 3 shows that, although the proposed ANN architecture can significantly outperform all of other prominent CA-based and DNN-based algorithms in the BETA dataset, the performance in the BETA dataset is still worse than that in the benchmark dataset. One of key reasons is the small number of calibration blocks in the BETA dataset. Fig. 6 shows that, as the number of calibration blocks decreases, the classification performance also decreases significantly. Therefore, more calibration blocks are still recommended for the proposed ANN architecture to achieve better recognition performance.

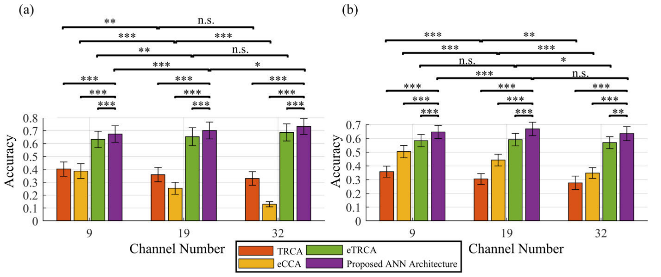

Fig. 7. Classification performance of the proposed ANN architecture and the prominent CA-based SSVEP recognition methods for different channel numbers. (a) Classification accuracy of 0.25s signal length for the benchmark dataset. (b) Classification accuracy of 0.4s signal length for the BETA dataset.

B. End-to-End and Global Optimization

Besides the compact structure, another key contribution of this study is that the advantages of the ANNs, i.e., the end-to-end and global optimization, is introduced to the individual SSVEP recognition. Compared to the conventional spatial filtering methods, the proposed method considers the variances within and between classes, and avoids the error accumulation, resulting in more optimal spatial filters and better performance as shown in Fig. 7. Because the signal qualities in different channels may be varied, superior calibration performance normally requires the suitable channel selection. As illustrated in Fig. 7, the CA-based methods calculate spatial filters without considering other model parameters and/or only considers the variances within classes, which obstruct the further performance improvements when the number of EEG channels increases. Nevertheless, owing to the end-to-end and global optimization, the proposed method can effectively avoid this obstacle. These results suggest that introducing the benefits of the ANN techniques is promising for overcoming the over channel selection issue in the EEG analysis. More discussions related to the advantages of fine-tuning sub-band weights and weights in the multi-head attention layer are illustrated in Sections S.V and S.VI of the supplementary material.

The global optimization of the proposed ANN architecture is achieved by the gradient descent, which requires lots of epochs to iteratively evaluate the loss function on all calibration signals in the training blocks. Compared to the training processes of the conventional spatial filtering methods that only solve the GEP shown in (1), the computational cost of the training process for the proposed method is much larger. However, after the entire training process, the SSVEP recognition process of the proposed ANN architecture is only based on the simple matrix operations, and thus does not need large computational cost. Therefore, the computational time of the proposed ANN architecture is much faster than that of some conventional spatial filtering methods that require to calculate the spatial filters of the new incoming EEG signals, such as the sCCA and the eCCA, as illustrated in Table II. On the other hand, our method achieves comparable running time with (e)TRCA, but it can provide significantly better accuracies and ITRs as shown in Fig. 3. In addition, since the trainable parameter number of the proposed ANN architecture is much smaller than the conventional spatial filtering methods as shown in Table I, the memory cost for storing the model of the proposed recognition is much smaller than that of the conventional spatial filtering methods. Moreover, due to the small size of the recognition model, the computational cost of the proposed method is much smaller than that of the DNN-based methods.

C. Limitations

Although the proposed ANN architecture can deliver the promising individual SSVEP recognition performance under the limited calibration data size, it is only at its infancy. There are still some practical problems. Firstly, kernels in the temporal filtering layer are pre-defined in this study, which are SSVEP reference and template signals. However, they are either too simple to present real EEG signals or easily affected by the external interference. According to the DNN-based SSVEP recognition methods [6], [22], [23], the recognition performance may be further improved by using calibration data to fine-tune the kernels in the temporal filtering layer. The prior knowledge of SSVEP signals, such as the dynamic model of SSVEP signals proposed in [43], and the transfer learning techniques of SSVEP signals introduced in [14], [26], [27], [28], and [29], may be helpful for designing the initial values and the constraints of optimizing kernels in the temporal filtering layer under the limited calibration data size. Secondly, most layers in the proposed ANN architecture are inspired by the unified framework of the spatial filtering algorithms and thus are limited by the linear operations. Incorporating more non-linear operations may be conducive to analyze nonlinear features of SSVEP signals mentioned in [44] and further improve the performance. Thirdly, the proposed ANN architecture does not adopt the new coming data to adjust trainable parameters, and thus cannot achieve the lifelong improvement. Therefore, it still requires enough calibration blocks as mentioned in Section IV-C. By utilizing the online adaption techniques such as the online adaptive CCA proposed in [45], the required calibration data size would be further reduced.

VI. Conclusion

This study proposes a task attention-based ANN architecture for the individual SSVEP recognition. The proposed task attention layer integrates the conventional spatial filtering algorithms into the ANN architecture. The conventional spatial filtering algorithms provide the prior knowledge of the SSVEP tasks, including the SSVEP reference/template signals, the SSVEP decoding schemes, and the common knowledge shared across stimuli. By incorporating the prior knowledge, the trainable parameter number in the ANN structure can be condensed to reduce the over-fitting effects under the limited calibration data size. Then, the individual SSVEP recognition can be benefited by the end-to-end and global optimization provided by the ANN technique. Simulation results on two public datasets show that, compared to prominent CA-based and DNN-based methods, the proposed ANN architecture can use the smallest trainable parameter number to provide the best performance in the individual SSVEP recognition. This research provides a promising future of using the prior knowledge to eliminate the number of redundant parameters in the SSVEP recognition models, and thereby introducing the benefits of the ANN techniques to the SSVEP analysis with less calibration.

References

[1] M. A. Lebedev and M. A. L. Nicolelis, "Brain–machine interfaces: Past, present and future," *Trends Neurosciences*, vol. 29, no. 9, pp. 536–546, Sep. 2006.

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

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

[4] C. M. Wong, B. Wang, Z. Wang, K. F. Lao, A. Rosa, and F. Wan, "Spatial filtering in SSVEP-based BCIs: Unified framework and new improvements," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 11, pp. 3057–3072, Nov. 2020.

[5] Y. Chen, C. Yang, X. Chen, Y. Wang, and X. Gao, "A novel training-free recognition method for SSVEP-based BCIs using dynamic window strategy," *J. Neural Eng.*, vol. 18, no. 3, Jun. 2021, Art. no. 036007.

[6] O. B. Guney, M. Oblokulov, and H. Ozkan, "A deep neural network for SSVEP-based brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 2, pp. 932–944, Feb. 2022.

[7] Z. Lin, C. Zhang, W. Wu, and X. Gao, "Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2610–2614, Dec. 2006.

[8] M. Nakanishi, Y. Wang, Y.-T. Wang, Y. Mitsukura, and T.-P. Jung, "A high-speed brain speller using steady-state visual evoked potentials," *Int. J. Neural Syst.*, vol. 24, no. 6, Sep. 2014, Art. no. 1450019.

[9] M. Nakanishi, Y. Wang, X. Chen, Y.-T. Wang, X. Gao, and T.-P. Jung, "Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 1, pp. 104–112, Jan. 2018.

[10] C. M. Wong et al., "Learning across multi-stimulus enhances target recognition methods in SSVEP-based BCIs," *J. Neural Eng.*, vol. 17, no. 1, Jan. 2020, Art. no. 016026.

[11] X. Chen, Y. Wang, S. Gao, T.-P. Jung, and X. Gao, "Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface," *J. Neural Eng.*, vol. 12, no. 4, Aug. 2015, Art. no. 046008.

[12] O. Friman, I. Volosyak, and A. Graser, "Multiple channel detection of steady-state visual evoked potentials for brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 4, pp. 742–750, Apr. 2007.

[13] Y. Zhang, P. Xu, K. Cheng, and D. Yao, "Multivariate synchronization index for frequency recognition of SSVEP-based brain–computer interface," *J. Neurosci. Methods*, vol. 221, pp. 32–40, Jan. 2014.

[14] C. M. Wong et al., "Inter- and intra-subject transfer reduces calibration effort for high-speed SSVEP-based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 10, pp. 2123–2135, Oct. 2020.

[15] G. Bin, X. Gao, Y. Wang, Y. Li, B. Hong, and S. Gao, "A high-speed BCI based on code modulation VEP," *J. Neural Eng.*, vol. 8, no. 2, Apr. 2011, Art. no. 025015.

[16] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung, and S. Gao, "High-speed spelling with a noninvasive brain–computer interface," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 44, pp. E6058–E6067, Nov. 2015.

[17] B. Liu, X. Chen, N. Shi, Y. Wang, S. Gao, and X. Gao, "Improving the performance of individually calibrated SSVEP-BCI by task-discriminant component analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1998–2007, 2021.

[18] N.-S. Kwak, K.-R. Müller, and S.-W. Lee, "A convolutional neural network for steady state visual evoked potential classification under ambulatory environment," *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0172578.

[19] J. Thomas, T. Maszczyk, N. Sinha, T. Kluge, and J. Dauwels, "Deep learning-based classification for brain–computer interfaces," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Banff, AB, Canada, Oct. 2017, pp. 234–239.

[20] N. Waytowich et al., "Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials," *J. Neural Eng.*, vol. 15, no. 6, Dec. 2018, Art. no. 066031.

[21] N. K. N. Aznan, S. Bonner, J. Connolly, N. Al Moubayed, and T. Breckon, "On the classification of SSVEP-based dry-EEG signals via convolutional neural networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Miyazaki, Japan, Oct. 2018, pp. 3726–3731.

[22] Y. Li, J. Xiang, and T. Kesavadas, "Convolutional correlation analysis for enhancing the performance of SSVEP-based brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2681–2690, Dec. 2020.

[23] A. Ravi, N. H. Beni, J. Manuel, and N. Jiang, "Comparing user-dependent and user-independent training of CNN for SSVEP BCI," *J. Neural Eng.*, vol. 17, no. 2, Apr. 2020, Art. no. 026028.

[24] K. Hagiwara and K. Fukumizu, "Relation between weight size and degree of over-fitting in neural network regression," *Neural Netw.*, vol. 21, no. 1, pp. 48–58, Jan. 2008.

[25] M. Bataineh and T. Marler, "Neural network for regression problems with reduced training sets," *Neural Netw.*, vol. 95, pp. 1–9, Nov. 2017.

[26] K.-J. Chiang, C.-S. Wei, M. Nakanishi, and T.-P. Jung, "Boosting template-based SSVEP decoding by cross-domain transfer learning," *J. Neural Eng.*, vol. 18, no. 1, Feb. 2021, Art. no. 016002.

[27] M. Nakanishi, Y. Wang, C. Wei, K. Chiang, and T. Jung, "Facilitating calibration in high-speed BCI spellers via leveraging cross-device shared latent responses," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 4, pp. 1105–1113, Apr. 2020.

[28] B. Liu, X. Chen, X. Li, Y. Wang, X. Gao, and S. Gao, "Align and pool for EEG headset domain adaptation (ALPHA) to facilitate dry electrode based SSVEP-BCI," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 2, pp. 795–806, Feb. 2022.

[29] Z. Wang, C. M. Wong, A. Rosa, T. Qian, T. Jung, and F. Wan, "Stimulus-stimulus transfer based on time-frequency-joint representation in SSVEP-based BCIs," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 2, pp. 603–615, Feb. 2023.

[30] X. Ma, S. Qiu, and H. He, "Time-distributed attention network for EEG-based motor imagery decoding from the same limb," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 496–508, 2022.

[31] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jul. 2017, pp. 6298–6306.

[32] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017.

[33] X. Dai et al., "Dynamic head: Unifying object detection heads with attentions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 7369–7378.

[34] Y. Liu, H. Zhang, D. Xu, and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," *Knowl.-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108146.

[35] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," 2021, *arXiv:2109.07270*.

[36] W. Tao et al., "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 382–393, Jan. 2023.

[37] D. Zhang, L. Yao, K. Chen, S. Wang, P. D. Haghighi, and C. Sullivan, "A graph-based hierarchical attention model for movement intention detection from EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2247–2253, Nov. 2019.

[38] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, and A. Etemad, "Classification of hand movements from EEG using a deep attention-based LSTM network," *IEEE Sensors J.*, vol. 20, no. 6, pp. 3113–3122, Mar. 2020.

[39] Z. Gao, X. Sun, M. Liu, W. Dang, C. Ma, and G. Chen, "Attention-based parallel multiscale convolutional neural network for visual evoked potentials EEG classification," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 2887–2894, Aug. 2021.

[40] D. Bahdanau et al., "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1409.0473

[41] Y. Wang, X. Chen, X. Gao, and S. Gao, "A benchmark dataset for SSVEP-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng* vol. 25, no. 10, pp. 1746–1752, Nov. 2017.

[42] B. Liu, X. Huang, Y. Wang, X. Chen, and X. Gao, "BETA: A large benchmark database toward SSVEP-BCI application," *Frontiers Neurosci.*, vol. 14, p. 627, Jun. 2020.

[43] S. Zhang, X. Han, X. Chen, Y. Wang, S. Gao, and X. Gao, "A study on dynamic model of steady-state visual evoked potentials," *J. Neural Eng.*, vol. 15, no. 4, Aug. 2018, Art. no. 046010.

[44] A. Notbohm, J. Kurths, and C. S. Herrmann, "Modification of brain oscillations via rhythmic light stimulation provides evidence for entrainment but not for superposition of event-related responses," *Frontiers Hum. Neurosci.*, vol. 10, p. 10, Feb. 2016.

[45] C. M. Wong et al., "Online adaptation boosts SSVEP-based BCI performance," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 6, pp. 2018–2028, Jun. 2022.