Enhancing Target Recognition Performance in SSVEP-Based Brain–Computer Interfaces via Deep Neural Networks with Pyramid Squeeze Attention

Xiao Wu, Ian Daly, Andrew Ty Lau,Weijie Chen, Chongfeng Wang , Andrzej Cichocki, *Life Fellow IEEE*, Jing Jin*, *Senior Member IEEE*

***Abstract*—Steady state visual evoked potential (SSVEP)-based brain-computer interfaces have been widely studied for their fast response speeds and high information transfer rates. However, how to fully utilize the potential information of existing subjects to realize the mining of common information among different subjects and then realize the information migration in a small amount of data scenarios is a difficult problem faced by current research. In order to solve the above problems, this study proposes a deep neural network based on the pyramid squeeze attention (PSA-DNN) mechanism to enhance the performance of SSVEP-BCI through common information migration. Specifically, the band-pass filtered EEG signals were first Fourier transformed to obtain the frequency domain information; subsequently, the frequency domain information is input into a deep neural network, followed by a spatial convolution step to extract spatial domain information. In order to further enhance the quality of information extraction, a pyramid attention module is introduced into the network to realize the enhancement of frequency domain and spatial domain information. Time domain information from the EEG signals is then mined using temporal convolution. Finally, the full connectivity layer is used to output the recognition results. The model is trained in a three-stage stepped approach for SSVEP target recognition. The first stage uses data from all participants in the training set for common information learning and transfers the model parameters trained in the first stage to the network model in the second stage. In the second stage, some of the information from participants in the test set is used for fine-tuning and to mine personalized information from these new participants. The third stage uses the remaining data from participants in the test set to produce classification results. The proposed method is systematically evaluated using the Benchmark and BETA datasets, where it demonstrates favorable performance compared to established baselines. These findings contribute theoretical insights and methodological references for the application of SSVEP-based brain–computer interfaces in real-world scenarios.**

***Index Terms*—Brain-computer interface, steady-state visual evoked potential, deep neural network, pyramid squeeze attention, target recognition.**

I. INTRODUCTION

Brain-computer interfacing (BCI) refers to a communication framework in which brain signals are directly translated into commands for external devices. It has a broad range of potential application prospects in the fields of medical rehabilitation, smart homes, gaming and entertainment, and intelligent education [1-5]. Currently, the three most popular BCI paradigms are based on motor imagery (MI) [6-10], event-related potentials (ERP) [11-13] and steady-state visual evoked potentials (SSVEP) [14-16].

The performance of SSVEP-based BCI systems is determined, in part, by the efficiency and accuracy of the decoding algorithms they use. A wide range of SSVEP decoding approaches has been introduced by researchers in recent years. Since the signals of SSVEP have similarities with sine-cosine signals, the target frequency of the SSVEP response can be identified by calculating the correlation between the EEG signal and a sine-cosine reference signal. Traditional decoding was based on the canonical correlation analysis (CCA) algorithm and over the years researchers have proposed a number of improvements to this base method [17]. For example, filter bank CCA (FBCCA) seeks to improve the effectiveness of the CCA algorithm by adding filter banks [18]. The extended CCA(ECCA) [19] and multiset CCA [20] also perform target frequency recognition by maximizing the correlation between the target signal and the reference signal in a training-free manner.

The above algorithms perform recognition based on a

This work was supported by Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project 2022ZD0208900 and National Natural Science Foundation of China under Grant 62176090; 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 JKH01241605 and key core technologies) under Grant BE2022064-1; in part by the Lingang Laboratory under Grant No.LGL8998.

Xiao Wu, Weijie Chen and Chongfeng 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,806237081@qq.com).

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).

Andrew Ty Lau is with the Shanghai Lansheng Brain Hospital Investment Co., Ltd, Shanghai 200336, China (e-mails: AndrewTyLau@126.com).

Andrzej Cichocki is with the Systems Research Institute of Polish Academy of Sciences, 01-447bWarsaw, and Nicolaus Copernicus University (UMK), 87-100 Torun, Poland (e-mail: cichockiand@gmail.com).

Jing Jin is 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, ECUST Medical-Engineering Integration Innovation Center (e-mail: jinjingat@gmail.com;) (Corresponding author: Jing Jin).

reference signal. However, different participants may produce different responses when subjected to target stimuli of different frequencies. Therefore, Nakanishi et al. proposed a decoding algorithm based on individual EEG template signals in an approach referred to as template signals-task-related component analysis (TRCA) [21]. By training the spatial filter to extract the spatial features of the SSVEP response and thus achieve the target classification of SSVEP, Wong et al. proposed a multistimulus extended version of CCA by optimizing the spatial filter through use of information in neighboring frequency bands. This method further enhances the effectiveness of the SSVEP decoding algorithm [22].

In recent years, deep learning methods have demonstrated superior feature mining capabilities in the fields of image segmentation [23,24], target detection [25], and video recognition [26,27], and thus have been gradually introduced into the decoding algorithms of SSVEP. Convolutional neural network (CNN) is one of the most important breakthroughs in the field of deep learning and is widely used in image processing tasks [28,29]. CNNs can automatically learn and extract features through different convolutional layers without the need to extract features manually. SSVEP is rich in nonlinear latent information and, compared with traditional methods, CNN is better equipped to deal with this information. CNN-based approaches outperform conventional techniques in multiple aspects, such as robustness to faults, autonomous feature learning, adaptive behavior, fine-grained feature resolution, and strong generalization ability.

Cecotti proposed a novel structure for convolutional neural networks [30]. The structure includes two hidden layers that switch the signal analysis inside the network from the time domain to the frequency domain. The method is an end-to-end data-driven approach that does not require any data preprocessing operations to achieve the task of classifying SSVEP signals. In order to further utilize the information of the SSVEP signals within different frequency bands, Zhao et al. proposed a filter bank convolutional neural network structure (FBCNN) for SSVEP classification [31]. The method preprocesses the SSVEP signals with band-pass filtering to obtain the signals in different frequency bands and then extracts and learns the harmonic features in the frequency bands through three parallel CNN channels to finally derive the correlation between the harmonics. The FBCNN method improves the performance of the CNN-based SSVEP classification method. A deep learning framework was developed by Guney et al. to capture spectral, spatial, and temporal characteristics of SSVEP signals collected from several electrode sites. The network applies convolutions over harmonic frequency bands, electrode arrays, and temporal windows, with fully connected layers used for classification, demonstrating promising results [32]. Liu et al. proposed a novel deep learning framework, EEGNetPSA, designed to capture subtle yet critical temporal dynamics in SSVEP signals, achieving strong performance under within-subject training conditions [33].

Some of the above traditional methods as well as neural network algorithms have achieved better classification scores in within-participant classification scenarios. However, in online SSVEP-based BCI system applications, different participants will be involved and the different responses of different participants to the SSVEP stimuli will lead to large differences in signals across participants. Therefore, it is important to seek to improve the classification performance of the model in cross-participant scenarios. In these scenarios, a classifier is first trained using data from a group of participants, and then tested using data from new, unseen, participants.

Waytowich et al. proposed a compact CNN network-EEGNet for classification of different participants [34]. Chen et al. proposed the SSVEPformer network structure to achieve SSVEP target recognition across participants [35]. Although both the above models achieved some success, they did not use the information about the new, unseen participants and ignored the personal characteristics of individual participants when performing cross-participant decoding.

In order to address the shortcomings of the above studies, Guney proposed a two-stage modeling strategy [32]. In this two-stage strategy, the training data from all participants were first used to train a global model, then the training data from individual participants were used to fine-tune the model, finally the test data from individual participants was used for testing to obtain the final classification results. This approach worked well, however, it still relied on a within-participant training strategy. If the global model trained in the first stage is applied to new participants, the performance of the model is greatly reduced. Zhang et al. proposed the ConsenNet framework, which aims to enhance SSVEP classification performance by leveraging information from existing participants [36]. However, during data augmentation, the method randomly selects only 10 participants, which does not fully exploit the underlying inter-participant correlations.

To solve the above problems, this paper proposes a deep neural network model based on the pyramid squeeze attention (PSA-DNN) to enhance the performance of SSVEP-BCI through common information migration. First, the mining of frequency domain information and spatial domain information is carried out through two-layer convolution, then pyramid squeeze attention is introduced to recalibrate the frequency domain information and spatial domain information, suppress redundant information, and enhance the weights of important information. Group convolution is then carried out to acquire the time domain information, and finally, the final recognition results are outputted through the fully-connected layer.

Our main innovations of this paper are as follows:

(1). We propose a novel SSVEP decoding architecture that directly exploits Fourier-transformed spectral components (real and imaginary parts) as model inputs. By jointly integrating frequency-domain convolution, spatial-domain convolution, and a pyramid squeeze attention mechanism for multi-scale feature recalibration-along with group convolution for complementary temporal modeling-the framework enables unified modeling of frequency, spatial, and temporal information. This design aligns closely with the neurophysiological characteristics of SSVEP signals and facilitates more discriminative feature representation.

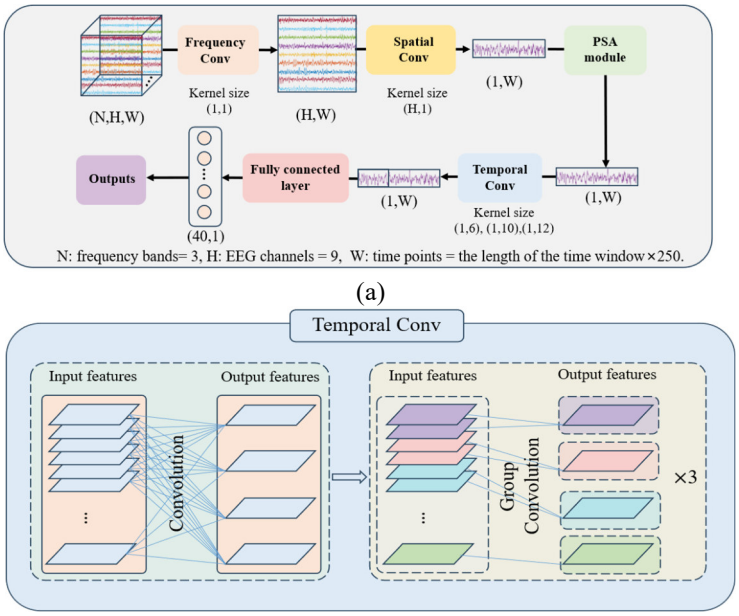


Fig. 1 (a) Structural framework diagram of PSA-DNN. The preprocessed signal first goes through the frequency domain convolution layer for frequency domain feature learning and extraction, and the second layer goes through spatial convolution for spatial feature learning and extraction. In order to further improve the effectiveness of frequency domain features and spatial features, the features are input into the PSA module, and the features are recali-fied, and then input into the time convolution layer for time feature extraction. The time-dependent information is captured, and finally the result is output through the fully connected layer to complete the classification task. **(b)** Specific structure of time-domain convolution. In the temporal convolution layer, three layers of group convolution are used to extract the time domain information. In addition, group convolution can greatly reduce the number of parameters and improve the computational efficiency.

TABLE I. The core information of the dataset.

Dataset	Subject	Block	Target	Channel	Stimulus frequency
Benchmark	35	6	40	64	8-15.8Hz
BETA	70	4	40	64	8-15.8Hz

(2). We develop a feature recalibration strategy tailored to SSVEP signals by embedding pyramid squeeze attention into the frequency-spatial processing pipeline. Unlike generic attention mechanisms, this design emphasizes stimulus-locked spectral components and spatial patterns associated with visual cortex activity, while suppressing redundant or non-task-related information, thereby improving robustness under low signal-to-noise conditions and short time windows.

(3). We introduce a structured three-stage training paradigm for cross-subject SSVEP decoding that explicitly balances global generalization and subject-specific adaptation. The approach first learns subject-invariant representations from large-scale multi-subject data, then performs lightweight fine-tuning using minimal calibration data from a new subject, and finally evaluates on unseen samples from that subject. This progressive strategy effectively mitigates inter-subject variability while reducing calibration burden.

The layout of this article is as follows: Section II introduces the datasets, network structures and experimental setup. Section III presents the experimental results. Section IV discusses the importance of each part of the proposed network structure and explores the influencing factors of the model performance. The final section includes a summary of the work and a discussion on the model's future prospects.

II. MATERIALS AND METHODS

A. Overview of Dataset Information

I. Benchmark dataset: The Benchmark dataset was collected from 35 participants. Each participant's data was recorded over 6 blocks, and each block contains 40 stimulus target presentation events (trials). The stimulus targets were presented to the participants on a computer screen arranged in a 5*8 matrix. The stimulus presentation frequency range was 8Hz-15.8Hz. The frequency interval was 0.2 Hz. The stimulation duration of each target was 6 s, and there was a 0.5 s cue before each target was presented. A total of 64 channels of EEG signals were recorded in each experiment, with a raw sampling rate of 1000 Hz which was down-sampled to 250 Hz.

II. BETA dataset: The BETA dataset is similar to the Benchmark dataset in that it also uses 40 stimulus targets with presentation frequencies ranging from 8 to 15.8 Hz, with intervals of 0.2 Hz. The BETA dataset was recorded from a larger number of participants (70 in total), and each participant's data was recorded over 4 blocks. The target stimulus durations used for participants 1 to 15 were 2 seconds,

and the stimulus durations of the remainder of the participants were 3 seconds. EEG data were recorded from 64 channels. The major difference from the Benchmark dataset is that the BETA dataset was acquired in a real environment and the signal-to-noise ratio of the signal is low.

The same preprocessing operation was used for both datasets. A Butterworth bandpass filter was used to divide the original signal into three frequency bands of 8-90Hz, 16-90Hz, and 24-90Hz. After filtering, the Fourier transform is performed on the data of each channel, and the real and imaginary parts of the transform results are extracted and spliced into the frequency band dimension respectively, so as to construct multi-band spectral features for model input.

B. Description of the model structure

In order to enhance the performance of SSVEP target recognition, a deep neural network model with a pyramid squeeze attention structure is proposed in this paper. The model contains six convolutional layers, a pyramid attention module and a fully connected layer (as shown in **Fig. 1(a)**). The specific structure of the model is as follows:

(1) The first part of the network module focuses on frequency domain feature mining of SSVEP signals. When a visual stimulus is flashed at a specific frequency, the SSVEP generated by the brain show a characteristic response related to the stimulus frequency and its harmonics. Notably, there are significant differences in the extent to which these harmonic components contribute to the SSVEP signal: typically, lower harmonic components such as the fundamental frequency (first harmonic) and the second harmonic have larger amplitudes, while the energy contribution of the harmonics tends to show a decreasing trend as the number of harmonics increases. This frequency domain feature distribution provides an important basis for frequency identification of SSVEP signals.

In terms of technical implementation, the module adopts a direct input strategy in the frequency domain, where the spectral features obtained from the original EEG signal after Fourier transform are used as inputs to the model. This processing avoids the noise interference in the time-domain signal and directly focuses on the effective information in the frequency domain. SSVEP responses have narrowband properties and are strictly locked to external stimuli, so maintaining fine frequency resolution is essential for effective feature extraction. If a large convolution kernel is used when processing the frequency domain, it may cause the blurring of fine-grained spectral features such as harmonics and intermodulation components, thereby weakening the discriminability between different targets. Based on this, a 1×1 convolution kernel is used in the frequency domain feature learning stage in this paper to ensure that the frequency resolution of the SSVEP signal and its physiological specificity are not lost. In addition, 1×1 convolution essentially performs frequency-based feature projection rather than spatial neighborhood aggregation, which is more suitable for frequency domain feature extraction.

In the specific design, we use a convolution kernel of size (1, 1) and configure 128 filters in the frequency domain branch for deep feature learning. This design has the following three advantages:

(1) Preserving the independence of frequency points: The 1×1 convolution acts on each frequency point in the form of dot product, which avoids unnecessary mixing between frequency neighborhoods and ensures that the harmonics and their high-order components are accurately preserved.

(2) Constructing a high-dimensional feature space: 128 filters provide the model with sufficient expressive power to capture the subtle differences between different frequency bands, harmonics, and intermodulation components;

(3) Adaptively learning frequency correlation: By automatically assigning weights to different harmonic components and spectral features, the network emphasizes informative frequency responses, leading to improved recognition performance.

In summary, the design of 1×1 convolution kernel not only conforms to the frequency characteristics of SSVEP signals, but also provides an efficient and physiologically reasonable modeling method for frequency domain feature projection, which helps to improve the frequency sensitivity and overall identification ability of the model.

(2) The second part of the model is used for mining spatial information. SSVEP signals are usually acquired synchronously by multi-lead electrodes, and the signals of each channel reflect the characteristics of neural electrical activities in different functional areas of the brain. In order to fully exploit the synergistic characterization ability among multiple channels, the second level of the network uses a spatial convolution operation to realize cross-channel feature integration.

The core objective of the spatial convolution operation is to analyze the spatial correlation between channels and reveal the functional coupling of different brain regions. During the data preprocessing phase, the choice of relevant electrodes needs to be consistent with the functional distribution of brain regions. Since SSVEP mainly originates from the posterior cortex of the brain, the design of the convolution kernel covering the entire visual-related area (a total of nine channels) matches the neurophysiological characteristics of the visual evoked potential, so as to achieve the effective capture of spatial coherence and ensure stable and physiologically valid spatial feature extraction. In contrast to the first layer, which focuses on frequency-domain feature extraction, this layer focuses more on the topological correlations brought about by the spatial distribution of electrodes. Through the dynamic weighted fusion of inter-channel signals, realized by the trainable convolutional kernel, the network is able to discover the contribution of each channel to spatial feature decoding and then extract discriminative spatial pattern features.

To implement this, the layer adopts a convolutional kernel structure of 9×1 dimensions, which is designed to retain the local detailed features of single-channel signals and realize the deep mining of cross-channel spatial correlation patterns. After processing in this layer, the network is able to effectively extract spatially discriminative feature representations from multi-channel signals, providing key spatial coding information

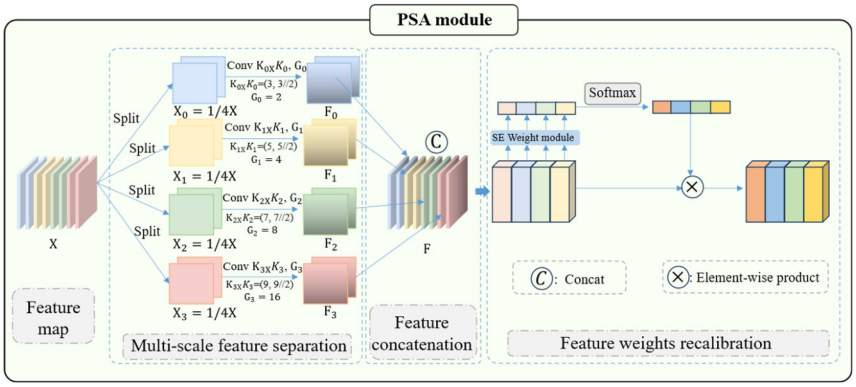


Fig. 2. Schematic diagram of the specific structure of the PSA module. First, the input features are separated through four parallel branches to realize the multi-view representation of the feature space. Subsequently, convolutional layers with different kernel sizes are introduced in each branch to learn feature weights at multiple scales. After the extraction, the features of each scale are concatenated to achieve the effective fusion of cross-scale information. Finally, the channel weights of the fused features were adaptively updated by the SE attention module, so as to obtain a more discriminative and effective feature representation.

for subsequent classification decisions. This feature extraction architecture significantly improves the network's ability to express spatial features, establishes a reliable spatial feature foundation for SSVEP signal decoding, and ensures optimal utilization of functional correlations between different channels.

(3) Design and implementation of the third part of the model, pyramid squeeze attention. After the first two modules complete the frequency-domain feature and spatial-domain feature extraction of SSVEP signals respectively, the network faces the challenges of feature redundancy and information interference. To solve this problem, this part innovatively introduces a multi-scale pyramid squeeze attention layer to optimize the feature characterization through feature recalibration [37]. The structure of this layer is shown in **Fig. 2**, and the module adopts a three-stage processing flow:

Multi-scale feature decomposition: A progressive down-sampling strategy is used to decompose the input feature map into four sub-feature spaces of different scales (original scale, 1/2 scale, 1/4 scale, and 1/8 scale), and 3, 5, 7, and 9 convolutions are applied to each scale space for local feature refinement. This hierarchical processing is able to capture both microscopic details and macroscopic modes, where the larger receptive field scale focuses on capturing global correlations across channels, while the fine scale preserves the harmonic details of the frequency domain features.

This paper introduces a multi-scale feature extraction architecture based on channel segmentation. The specific implementation process is as follows: First, the input feature map X is divided into channel dimensions and decomposed into S sub-feature maps, which are denoted as $[X_0, X_1, \ldots, X_{s-1}]$. where the channel dimension of each sub-feature map is:

$$C' = \frac{C}{s} \tag{1}$$

The sub-feature maps obtained from each segmentation can be represented as:

$$X_i \in R^{C' \times H \times W}, i = 0,1, \ldots, S-1 \tag{2}$$

Here, C represents the number of output channels = 128, H represents the number of EEG channels = 9, and W represents the number of time points = the length of the time window× 250.

In order to fully extract multi-scale spatial features, this method uses an adaptive grouping convolution operation for each sub-feature map. Through experimental validation, we find the following optimal relationship between the convolution kernel size K and the number of groupings G:

$$G = 2^{\frac{K-1}{2}} \tag{3}$$

The specific feature extraction process can be expressed as follows:

$$F_i = Conv(K_i \times K_i, G_i)(X_i), i = 0,1,2, \ldots, S-1, \tag{4}$$

where each parameter is satisfied:

$$K_i = 2 \times (i+1) + 1, G_i = 2^{\frac{K_i-1}{2}}, F_i \in R^{C' \times H \times W} \tag{5}$$

Ultimately, the fusion of multi-scale features is achieved through channel splicing operations:

$$F = Cat([F_0, F_1, \ldots, F_{S-1}]) \tag{6}$$

In order to establish the competitive relationship between scales, the softmax function is used to normalize the attention weights and obtain the attention distribution with interpretability:

$$att_i = Softmax(Z_i) = \frac{exp(Z_i)}{\sum_{i=0}^{S-1} exp(Z_i)} \tag{7}$$

Channel dimension modulation of raw features based on normalized attention coefficients is performed:

$$Y_i = F_i \odot att_i, i = 1,2,3, \dots S-1, \quad (8)$$

where $\odot$ denotes a channel-by-channel multiplication operation. The modulated multiscale features are finally integrated along the channel dimension to output enhanced features that incorporate multiscale contextual information:

$$Out = Cat([Y_0, Y_1, \dots, Y_{S-1}]) \quad (9)$$

This mechanism achieves synergistic optimization of local and global features through a hierarchical attention design. Softmax normalization establishes inter-scale competition and the feature modulation process preserves the unique features of each scale.

Cross-scale feature fusion: After unifying the features of each scale to the original resolution by bilinear interpolation, a fusion strategy combining channel concatenation and 1×1 convolution is adopted. This design not only retains the unique characterization advantages of each scale, but also establishes cross-scale feature correlation, especially enhancing the synergistic expression of fundamental frequency harmonic features and spatial topology features.

Dynamic Feature Re-weighting: An improved Squeeze-and-Excitation mechanism is introduced, which first generates channel-level statistical descriptors through global average pooling, then learns the dependencies of each feature channel using a two-layer fully connected network before, finally, generating adaptive weights through Sigmoid activation. For the input set F of multi-scale feature maps, a set of corresponding channel attention weights are generated for each scale feature map by the Squeeze Excitation (SE) module:

$$Z_i = SEWeight(F_i), i = 0,1,2, \dots, S-1, \quad (10)$$

where each $Z_i \in R^{C' \times 1 \times 1}$ denotes the channel importance distribution of features at that scale. The attention weights extracted from each scale are subsequently spliced and integrated to construct the global attention representation:

$$Z = Z_0 \oplus Z_1 \oplus \dots \oplus Z_{S-1}, \quad (11)$$

enabling attentional information interactions across scales.

The process pays special attention to the enhancement of discriminative features suitable for SSVEP decoding (e.g., frequency bands corresponding to dominant harmonics), while suppressing redundant features.

(4) The fourth part is used for mining time-domain information. As a typical time series, SSVEP signals contain critical temporal dynamic information. This module deeply mines the time-varying properties of EEG activities through a multi-scale convolutional architecture to accurately capture the neural response patterns that are phase-locked with visual stimuli. The parallel processing strategy of convolutional kernels with differentiated sizes is adopted to realize the synergistic extraction of transient and steady-state time-domain features, which significantly improves the richness of feature expressions.

Multi-scale convolutional kernels work in concert to form a key mechanism for time-domain feature extraction. The fine-scale convolutional kernels specialize in signal micro-features, which can sensitively detect transient events such as amplitude bursts and short-duration oscillations, and effectively capture the fast neural response at the initial stage of SSVEP stimulation. This high temporal resolution design significantly enhances the accuracy of recognizing dynamic patterns at the millisecond scale. The macroscale convolutional kernel, on the other hand, focuses on the overall signal characteristics and extracts the periodic patterns of steady-state oscillations through wide time window analysis, so as to accurately resolve the sustained rhythmic patterns of SSVEP signals that are phase-locked to the stimulus frequency. This complementary design enables the network to recognize both fast transient responses and capture stable time-domain oscillatory features.

The small convolution kernels of (1,4) and (1,6) are used to focus on the local details of the signal, while the macro convolution kernels of (1,10) and (1,12) are applied to analyze the overall waveform characteristics (As shown in **Fig. 1(b)**). This design realizes the multilevel analysis of the time-domain signal: the fine convolution kernel captures the millisecond transient response, and the macroscopic convolution kernel extracts the periodic steady state pattern. The synergistic work of different receptive fields enables the network to perceive both the microscopic fluctuations of the signal and grasp its macroscopic evolutionary patterns, constructing a complete time-domain feature space for making classification decisions.

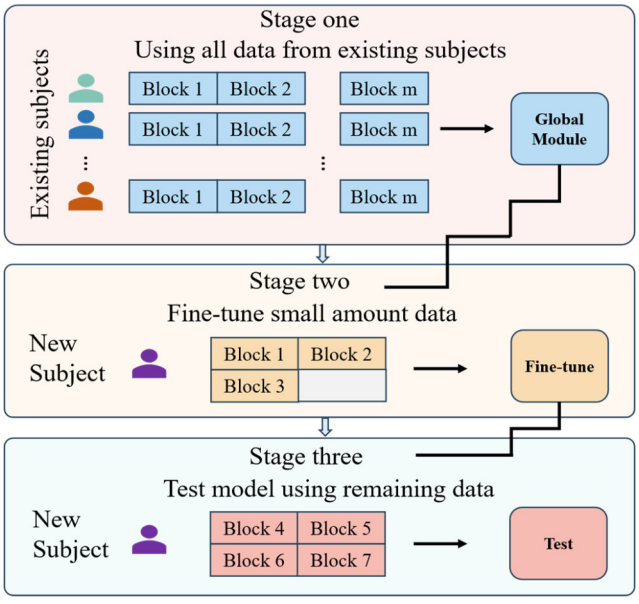


Fig. 3. A three-stage stepped model training approach. A subject independent global model is trained using all available subjects except the target test subject. For the benchmark dataset, this corresponds to the classical leave-one-subject (LOSO) strategy: in each fold, 34 subjects are used for training and 1 subject is taken out as the new (unseen) subject. For the BETA dataset, 69 subjects were used for training and 1 subject was used as a new subject. This phase enables

the model to learn generic, cross-topic SSVEP representations. Stage 2 - Topic-specific fine-tuning (learning personalized patterns). From the retained "new subjects", only a small fraction of trials (e.g., one or two blocks, depending on the dataset configuration) are used to fit the global model. This step allows the model to capture subject-specific spectral-spatial variations that are known to be important in EEG decoding. Stage 3 - Final evaluation of the remaining data for the new subjects.

The synergistic effect of multi-scale convolutional kernels aims to significantly improves the expression ability of time domain features. By fusing local details and global trend features, the network realizes three-dimensional analysis of the EEG signals spatio-temporal characteristics, in order to attempt to enhance the discriminative ability of the model. The feature refinement module optimizes the feature representation through the convolution operation, highlighting informative time domain patterns while suppressing background noise.

C. Model Training Strategy

In this paper, a progressive model training strategy consisting of three stages is proposed (the overall process is shown in **Fig. 3**). The core goal of this strategy is to achieve the efficient integration of "general feature learning" and "personalized adaptation", so as to balance the cross-subject generalization ability and individual level recognition performance of the model. In the first stage (global model training stage), we integrate the complete data set of all subjects to learn stable and consistent neural response patterns across different subjects through large-scale cross-individual data, and then construct a basic model with strong generalization ability. In this stage, a leave-one-subject-out (LOSO) strategy is adopted. For the Benchmark dataset, the data of 34 subjects are used for global model training. For the BETA dataset, the data of 69 subjects were used for model training. Through this process, the model is able to extract the SSVEP feature representation shared across subjects, which provides a solid and structured feature foundation for the subsequent personalized fine-tuning stage.

The second phase (personalized fine-tuning phase) employs a transfer learning paradigm with a global model as initialization and adaptive tuning of model parameters using limited calibration data from targeted new subjects. This critical step enables the model to gradually adapt to target subject-specific EEG feature variations, such as individual-specific brain anatomical structure differences, electrode impedance properties, etc., while retaining common features through a small-sample learning mechanism.

To rigorously evaluate the model, the final testing employs a hold-out strategy, applying it to data from new participants who had no exposure to the fine-tuning phase. This data partition method effectively avoids data leakage and ensures the objectivity and reliability of the model evaluation results.

D. Methods of comparison

We compared the performance of our proposed model to seven state of the art methods, including three traditional algorithms and four deep learning methods:

(1) ECCA [19]: Separate calibration data and artificial sinusoidal signals were used.

(2) iFuzzyTL [38]: iFuzzyTL integrates fuzzy inference systems with attention mechanisms to optimize input signal processing and classification, presenting results in a format readily comprehensible to humans.

(3) DB-AFDNet [39]: The Dual-Branch Attention Mechanism Frequency Domain Network (DB-AFDNet) is capable of independently decoding the real and imaginary spectral components, aiming to acquire more discriminative and generalisable features.

(4) DNN [32]: Deep Neural Network (DNN) realizes adaptive extraction and learning of dense features by stacking multilayer convolutional structures.

(5) EEGNetPSA [33]: Using EEGNet as the base network, a PSA module is subsequently integrated to accomplish the SSVEP target classification task.

(6) SSVEPformer [35]: SSVEPformer adopts the Transformer structure to characterise SSVEP signals.

(7) ConsenNet [36]: This teacher-student network model is used to learn and characterize the diverse features of the SSVEP.

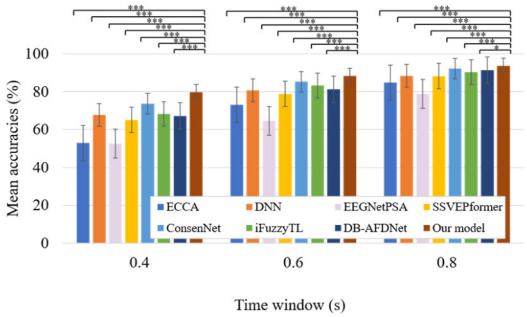
III. RESULTS

A. Detailed setup and performance evaluation

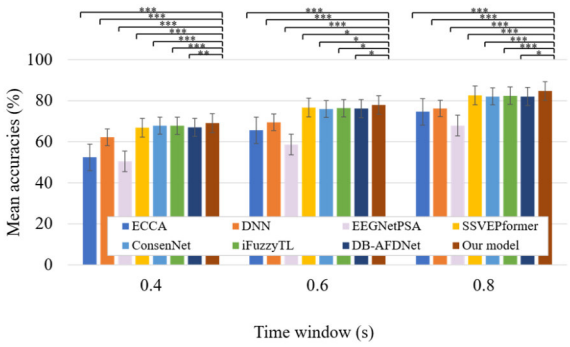
All deep learning models in this study (including our proposed method) are implemented based on PyTorch. The training process uses the Adam optimizer with an initial learning rate set to 1e-4 and a batch size of 128. Model training is divided into two phases: 80 rounds of global training and 300 rounds of fine-tuning. An early stopping mechanism (50 rounds are terminated if the validation loss does not improve) is used to prevent overfitting and improve training efficiency. The model optimization objective is to minimize the cross-entropy loss [40].

$$\frac{1}{D_b}\sum_{i=1}^{D_b} -\log\left(s_i(y_i)\right) + \lambda|w|^2 \qquad (12)$$

The parameters are updated using a stochastic gradient descent algorithm, and L2 regularization (with a penalty factor of 1e-10) is introduced to further normalize the model parameters. To ensure fairness in the comparison, all models in this paper use a uniform preprocessing procedure and training configuration.



(a)



(b)

Fig. 4. (a) On the Benchmark data set, 0.4 s, 0.6 s and 0.8 s time windows are used, and the experimental results of the proposed method and other comparison methods are compared. (b) On the Beta data set, 0.4 s, 0.6 s and 0.8 s time Windows are used, and the experimental results of the proposed method and other comparison methods are compared. The symbols *, **, and *** represent significance levels of $p<0.05$, $p<0.01$, and $p<0.001$, respectively.

In this study, classification accuracy and information transfer rate (ITR) are used as performance evaluation metrics. The accuracy rate is calculated as the ratio of the number of correctly predicted samples to the total number of tested samples and reflects the overall classification ability of the model. ITR is used to evaluate the communication efficiency of the BCI system and is calculated as follows [41]:

$$ITR = \frac{60}{T_W}\left[log_2 N + PLog_2 P + (1-P)log_2 \frac{1-P}{N-1}\right] \quad (13)$$

where P denotes the accuracy and N denotes the number of targets.

B. Performance Evaluation on Benchmark Dataset

We evaluated the performance of the models using the Benchmark dataset, which contains SSVEP data from 35 participants. The experiments were conducted using a leave-one-out cross-validation strategy: the global model was trained using data from 34 participants in the first phase, and individualized fine-tuning was performed in the second phase using a small amount of calibration data from the remaining participants. Specifically, fine-tuning was performed using 3 blocks of data from participants in the test set. The third phase tested the model using the remaining unseen data from the participants in the test set.

To fully evaluate the method performance, we used three time windows of 0.4 seconds, 0.6 seconds and 0.8 seconds for analysis. The experimental results are shown in **Fig. 4(a)**, which indicate that under the most challenging 0.4-second short time window condition, our proposed method demonstrates a significant performance advantage: the recognition accuracy is improved by 27.12%, 14.7%, 11.45%, and 6.0%, respectively, compared to the three mainstream methods, namely EEGNetPSA, SSVEPformer, iFuzzyTL, and ConsenNet. Under the condition of 0.6 seconds medium duration, there is still a 3.1% performance improvement relative to ConsenNet. It is worth noting that our proposed method still achieves the best average result even in the case of the 0.8-second long time window.

After a comprehensive comparison of the seven existing most advanced methods (excluding the algorithm proposed in this paper), we found that this model achieved the highest decoding performance under all time window conditions. Further statistical significance analysis indicated that the performance improvement of this method was statistically significant. The above results demonstrate that the proposed algorithm shows stable and significant advantages in different application scenarios and time window settings.

C. Performance Evaluation on BETA Dataset

To further evaluate the generalization ability and robustness of our proposed method we conducted supplementary experiments on the more challenging BETA dataset. This dataset contains SSVEP recordings from 70 participants and, uniquely, the data collection environment is an outdoor scene. The signal quality is significantly lower than that of the Benchmark dataset, which was recorded under laboratory conditions. This evaluation, using data recorded under more challenging scenarios, better reflects the practical application value of our proposed method.

For this experiment also used three typical time windows of 0.4 s, 0.6 s and 0.8 s duration for system evaluation, and the fine-tuning also used three blocks of data from participants in the test set. **Fig. 4(b)** indicates that the proposed approach continues to perform favorably under a brief 0.4-second time window, despite the reduced signal-to-noise ratio. Specifically, the accuracy is improved by 2.4%, 1.28%, 2.09% and 1.3% compared to SSVEPformer, iFuzzyTL, DB-AFDNet and ConsenNet, respectively.

The performance advantage of 11.44 percentage points over the iFuzzyTL is particularly noteworthy. This suggests that our proposed method is particularly robust to noisy environments. A stable advantage of 1.9% is maintained relative to ConsenNet in the 0.6 s medium duration condition. Even under the 0.8-second long time window condition where the signal quality is limited, our proposed method still shows some performance advantages. These results demonstrate that our proposed method maintains a stable performance advantage in both controlled laboratory environments (Benchmark dataset) and more challenging real-world scenarios (BETA dataset), showing excellent generalization ability and environmental adaptability.

It is worth noting that to further validate the technical advantages of this work, we conducted a direct comparison with EEGNetPSA using the identical dataset and training strategy. Experimental results demonstrate that the proposed frequency-spatial-temporal dedicated architecture significantly outperforms EEGNetPSA across key performance metrics: in cross subject validation tasks, the average classification accuracy improved by 18.6% under 0.4 seconds.

The core reason for this performance improvement lies in our architecture's tailored design for EEG signals' non-stationary and multi-band characteristics. Specifically, we designed specialized frequency convolution layers and multi-scale temporal convolution layers, whereas EEGNetPSA employs a generalized deep separable convolution architecture that lacks this specificity. Furthermore, the PSA module is positioned after spatio-temporal feature extraction in this work, serving as a refinement bottleneck

Table II. Classification accuracies (%) and standard deviation (%) of models fine-tuned using different calibration data.

Block	Benchmark			BETA		
	0.4 s	0.6 s	0.8 s	0.4 s	0.6 s	0.8 s
1	72.01±18.37	82.73±15.73	90.63±11.58	60.09±18.75	70.65±17.16	77.43±15.67
2	77.06±17.61	86.45±14.26	92.55±10.45	65.81±18.40	75.40±16.36	81.13±14.18
3	79.70±17.02	88.30±13.23	93.56±9.28	69.10±19.01	77.96±15.29	84.78±13.09

Table III. ITRs (bits/min) and standard deviation (bits/min) of models fine-tuned using different calibration data.

Block	Benchmark			BETA		
	0.4 s	0.6 s	0.8 s	0.4 s	0.6 s	0.8 s
1	207.21±74.82	210.36±57.84	206.09±39.81	157.04±72.17	163.91±59.10	159.78±48.26
2	230.14±74.12	225.60±54.29	213.64±36.87	180.36±73.44	181.43±58.08	171.84±45.23
3	242.75±73.16	233.68±51.72	217.52±33.71	194.18±76.27	191.59±55.84	182.56±42.56

Table IV. Accuraces (%) and standard deviation (%) of ablation experiments in four different situations.

Case	Benchmark			BETA		
	0.4 s	0.6 s	0.8 s	0.4 s	0.6 s	0.8 s
w/o spatial FFT&PSA	77.91±17.54	86.78±14.08	92.57±10.35	67.91±18.10	75.70±16.61	81.04±14.60
with PSA	79.54±16.91	87.91±13.39	93.36±9.50	68.58±18.27	77.37±15.68	82.29±13.33
with FFT	78.65±17.23	87.43±13.74	92.96±9.88	68.50±18.36	77.08±15.99	81.96±13.94
our model	79.70±17.02	88.30±13.23	93.56±9.28	69.10±19.01	77.96±15.29	84.78±13.09

Table V. ITRs (bits/min) and standard deviation (bits/min) of ablation experiments in four different situations.

Case	Benchmark			BETA		
	0.4 s	0.6 s	0.8 s	0.4 s	0.6 s	0.8 s
w/o spatial FFT&PSA	234.31±73.79	227.15±53.73	213.78±36.49	189.94±73.52	183.47±58.92	172.44±45.99
with PSA	241.83±72.50	231.87±51.65	216.77±34.22	193.05±74.44	189.47±56.27	176.24±55.84
with FFT	237.67±73.22	229.88±52.44	215.26±35.37	192.74±74.92	188.41±57.51	175.33±44.38
our model	242.75±73.16	233.68±51.72	217.52±33.71	194.18±76.27	191.59±55.84	182.56±42.56

for highly discriminative features. In contrast, EEGNetPSA embeds PSA as an independent module within the convolutional stack, resulting in significant differences in its adaptability and contribution to EEG decoding tasks.

IV. DISCUSSION

A. A feasible approach to categorizing scenarios between participants for SSVEP data

The large target instruction sets available in SSVEP-based BCI systems support their deployment in a broad range of practical scenarios. Despite numerous research results, target recognition of EEG signals, especially accurate decoding for multi-target scenarios, still faces significant challenges. Existing methods generally suffer from two key limitations: one, most algorithms only achieve good results on participant-specific data, with a significant performance degradation when faced with new participants. This stems from the failure of the models to adequately learn common features across participants. Second, existing methods tend to neglect the important value of individual-specific information.

To address these issues, this study innovatively proposes a three-phase training and testing paradigm, which achieves a balance between model generalization and specificity by organically combining the learning of common features and the adaptation of individualized features. Specifically, the first stage focuses on cross-participant common feature extraction, the second stage fine-tunes the model for individual specificity, and the third stage applies the model to new participants. This incremental learning strategy not only captures stable neural response patterns at the population level, but also preserves the unique EEG features of individuals, which significantly improves the system's adaptability to new users.

Table VI. Classification Accuracies (%) of the PSA-DNN after removing the spatial convolution layer and the temporal convolution layer.

Case	Benchmark			BETA		
	0.4 s	0.6 s	0.8 s	0.4 s	0.6 s	0.8 s
w/o spatial	57.25±20.50	68.76±19.00	79.86±17.40	51.29±19.39	59.61±19.37	66.28±19.23
w/o temporal	74.13±17.53	83.71±15.11	90.94±11.36	63.80±18.42	72.81±17.07	78.19±15.23
PSA-DNN	79.70±17.02	88.3±13.23	93.56±9.28	69.10±19.01	77.96±15.29	84.78±13.09

Table VII. The ITRs (bits/min) of the PSA-DNN when the spatial convolutional layer and the temporal convolutional layer are removed.

Case	Benchmark			BETA		
	0.4 s	0.6 s	0.8 s	0.4 s	0.6 s	0.8 s
w/o spatial	147.72±75.61	158.53±62.47	168.99±52.44	125.43±67.90	128.36±59.72	127.69±53.46
w/o temporal	216.30±71.87	214.33±56.02	207.37±38.94	172.97±71.67	172.84±59.20	162.97±46.87
PSA-DNN	242.75±73.16	233.68±51.72	217.52±33.71	194.18±76.27	191.59±55.84	182.56±42.56

Table VIII. The computational complexity of different models under 0.4 s.

Model	Params	First Stage Training Time (s)	Second Stage Training Time (s)	Inference Time per Sample (ms)	Accuracy (%)
SSVEPformer	2614276	50.75257	17.30714	0.308286	47.33
EEGNetPSA	258000	59.43000	7.72500	0.093370	52.58
DNN	413926	44.14286	15.33714	0.203429	57.71
ConsenNet	1025876	1020.15000	660.25690	0.361256	73.62
Our model	872528	187.50910	67.58457	0.652458	79.82

B. Impact of calibration data volume on model performance

In this study, an incremental three-stage training strategy was used. In the second stage the model was fine-tuning with calibration data from new participants. To systematically assess the effect of the amount of calibration data on model performance, we designed a comparative experimental scheme. As shown in **Table II** and **Table III**, the performance changes under three calibration data volume conditions examined over the two datasets: 1 block, 2 blocks, and 3 blocks.

The experimental results show two important findings: first, the model performance is significantly and positively correlated with the amount of calibration data. As the amount of calibration data increases from 1 block to 3 blocks, the recognition accuracy continues to improve, suggesting that richer individual data helps the model learn more discriminative feature representations. This phenomenon was particularly prominent in the most challenging 0.4-second short time window condition.

Second, the information transfer rate (ITR) showed an increasing and then decreasing trend, which reflects the trade-off between time window length and system efficiency. Under all test conditions, fine-tuning using 3 blocks of data resulted in optimal performance, a result that provides an important reference for calibration time settings in practical applications. These findings confirm that appropriately increasing the amount of calibration data can effectively improve the model's adaptability to new users.

C. Ablation experiments

In this study, the contribution mechanism of FFT frequency domain preprocessing and Pyramid Attention Module (PSA) to the model performance is analyzed through a set of ablation experiments. As shown in **Table IV** and **Table V**, the experiments are designed with four progressive comparison schemes: the base model (Case 1, without FFT and PSA), the PSA-enhanced model (Case 2), the FFT preprocessing model (Case 3), and the full model (Case 4, with both FFT and PSA integrated). By analyzing the experimental results from multiple perspectives, we obtain the following important findings:

1) The effectiveness of the attention mechanism of the PSA module: case 2 shows significant performance improvement compared to case 1 under all test conditions. This confirms the key role of the pyramidal attention structure in feature selection and cross-scale information fusion. Attentional weight visualization analysis showed that the PSA module was able to automatically focus on the brain regions and frequency bands most relevant to the stimulus frequency.

2) The necessity of FFT preprocessing: Case 3 exhibits a steady performance improvement over Case 1. This advantage is more obvious in low signal-to-noise ratio conditions. Frequency domain analysis shows that FFT conversion effectively enhances the signal-to-noise ratio of fundamental and harmonic components, providing a better input representation for subsequent feature extraction.

3) Synergistic enhancement effect between modules: Case 4 achieves optimal performance in all test scenarios, and the ITR metrics are synchronized with significant improvement (up to 35 bits/min). This synergistic effect stems from the organic combination of the high-quality frequency-domain features provided by the FFT and the feature selection capability of the PSA, forming a complementary enhancement mechanism.

It is worth noting that this performance advantage is particularly prominent under the long-time window (0.8 s) condition, and the improvement in performance of the full model over the base model reaches 3.7% under the outdoor acquisition condition of the BETA dataset. These findings not only provide a solid theoretical basis for the model design, but also indicate an important direction for future optimization of SSVEP-BCI systems.

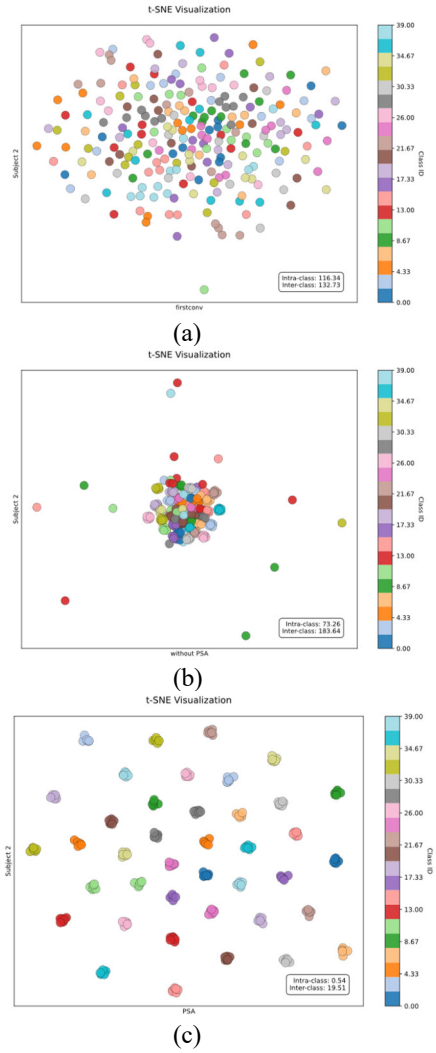


Fig. 5.(a) Feature distribution of the SSVEP signal after convolution of the first network layer. (b) Distribution of features in the absence of PSA modules. (c) Distribution of features using the PSA module.

In addition, we further investigate the role of spatial convolution versus temporal convolution in the model. To this end, we performed ablation experiments in two datasets with different time windows by removing spatial and temporal convolutional layers, respectively. **Tables VI and VII** show the classification accuracy and ITR of the model under different settings, respectively. Here, 'w/o spatial' means removing spatial convolutional layers, and 'w/o temporal' means removing temporal convolutional layers. From the experimental results, we can see that removing the temporal convolutional layer results in about 2.62-6.59% performance degradation of the model, while removing the spatial convolutional layer results in a more significant performance degradation of about 13.78%-22.45%.

The results indicate that during the multi-domain feature extraction process of the model, the contribution of spatial convolution is more prominent compared to temporal convolution, suggesting that spatial domain information plays a more significant role in characterizing the neural response patterns of SSVEP. However, although the importance of spatial convolution is more significant, temporal convolution also plays an indispensable role in capturing rhythmic neural dynamics. Overall, spatial convolution and temporal convolution form a complementary and collaborative feature learning mechanism in the network, and their combined effect is crucial for fully exploring the separable features of SSVEP and significantly improving the overall recognition performance of the model.

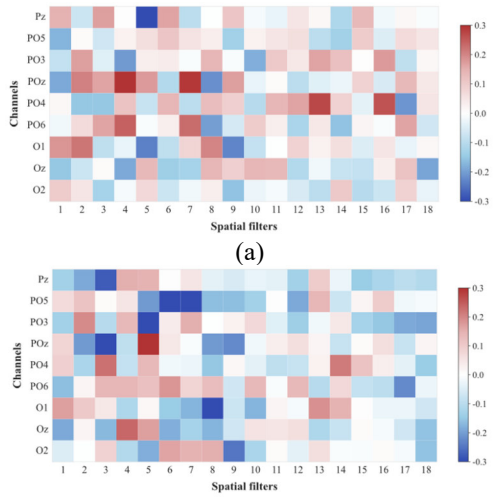
D. Model Visualization

To explore the progression of feature extraction, t-SNE [42] was employed for dimensionality reduction and visualization, with the corresponding results shown in **Fig. 5**. By tracking the evolutionary trajectories of the features in each layer of the model, the gradual improvement process of the representation quality can be clearly observed. In order to quantitatively assess the feature separability, we introduced the inter-class Euclidean distance as an evaluation index and conducted a comparative analysis of the feature distributions over the three participants. The experiment was set up with four comparison groups: the baseline model (no FFT and PSA), the FFT-enhanced model, the PSA-enhanced model, and the full model (FFT and PSA).

The visualization analysis reveals three important findings: first, the baseline model has the highest feature aggregation (mean inter-class distance = 132.73), resulting in blurred classification boundaries; second, the introduction of the FFT or PSA module alone results in an obvious trend of separation of the feature distributions (the inter-class distance is enhanced to 183.64 and 19.51, respectively), which corresponds to an increase in accuracy of about 3% ; most importantly, the full model exhibits an optimal feature distribution pattern (interclass distance = 0.54), with different classes of features forming a clear clustering structure. The above results indicate that the PSA module plays a crucial role in improving the model performance. By separating and re-aggregating the features, and achieving adaptive feature calibration, this module can guide the network to learn more discriminative features, thereby significantly enhancing the classification recognition performance. This visual evidence confirms the superiority of multi-module synergy from a geometric perspective and provides an intuitive explanation for the effectiveness of the model.

To further analyze the model's feature attention pattern in the spatial domain, we visualized the channel weights of the trained model across the nine electrodes. Specifically, we selected two representative subjects in the dataset and presented their channel weights in the form of heat maps, as shown in **Fig. 6**. It can be observed from the figure that POz and Oz electrodes show significantly higher weight values, and the high-weight regions are mainly distributed in these two electrodes and their surrounding locations. This result is highly consistent with previous studies: POz and Oz are located in the corresponding regions of the occipital visual cortex and are the most prominent and stable response channels for SSVEP signals [43]. The model automatically learned higher weights on these electrodes,

indicating that it successfully captured the neurophysiological characteristics of SSVEP. This phenomenon further demonstrates the capability of the proposed approach to learn spatial features, as it enables the model to selectively attend to brain regions with more pronounced neural activity, thereby enhancing recognition accuracy.



(a)

(b)

Fig. 6. Channel weight distributions after the spatial convolution layer for two representative subjects ((a) subject 4, (b) subject 13) in the Benchmark dataset. The visualization results are based on the classification scenes of three training blocks, all with a data length of 0.6 s. Each column in the figure corresponds to the channel weight extracted by a spatial filter, while each row represents the weight change of each electrode under all spatial filters.

E. Computational complexity

We selected several representative baseline models for comparison with the proposed method, with specific experimental results presented in **Table VIII**. Regarding model parameter scale, although our approach does not possess the minimal number of parameters, its parameter count is markedly lower than that of SSVEPformer and ConsenNets. Simultaneously, while maintaining a low parameter count, this method achieved the highest classification accuracy, demonstrating superior performance-complexity trade-off capabilities. In computational efficiency, our model demonstrates high performance in both training and testing times. Notably, during testing, its inference time falls well below the 0.4-second stimulus duration, rendering it negligible for real-time systems. Overall, our approach achieves a favourable trade-off between accuracy, model complexity, and computational efficiency, fully validating its practicality and convenience for real-world applications. We selected several baseline models for comparison with our proposed approach, with detailed results presented in **Table VIII**. Regarding model parameter count, while our model is not the most minimal, it remains comparatively sparse relative to SSVEPformer and ConsenNets, whilst achieving the highest accuracy. Both training and testing times are notably brief, rendering them negligible compared to the 0.4-second stimulus duration. These results demonstrate the practicality and convenience of our proposed method.

F. Limitations and Future Work

Although our proposed method demonstrates notable improvements in SSVEP-based brain-computer interface (BCI) target recognition, several limitations remain and warrant further investigation.

(1) Cross-dataset generalization remains unexplored. We have evaluated our proposed model on two large-scale SSVEP datasets; however, we have not assessed cross-dataset performance. Future work will investigate cross-dataset training strategies to more thoroughly evaluate the robustness and generalization of the model in varied recording environments.

(2) Computational cost limits real-time deployment. Despite achieving high recognition accuracy, the multi-band FFT processing and PSA module introduces additional computational overhead, which may restrict deployment in low-latency BCI applications. In future studies, we plan to develop lightweight model architectures to reduce computation cost, accelerate inference speed, and support efficient online BCI implementations.

(3) Subject-specific calibration is still required. The current three-stage training pipeline still relies on a small amount of calibration data for new subjects. Minimizing or eliminating this requirement remains an open challenge in BCI research. Future work will explore advanced subject adaptation and transfer learning techniques to reduce the dependency on subject-specific calibration.

(4) Neurophysiological interpretability requires deeper validation. Although spatial- and frequency-domain saliency analyses are provided, more in-depth neurophysiological validation (e.g., source localization) could further enhance the biological interpretability and neuroscientific relevance of our proposed model. In future research we will integrate such analyses to strengthen theoretical insights into SSVEP generation mechanisms.

V. CONCLUSION

To enhance decoding performance in SSVEP-BCI applications, this work proposes a deep neural network model that leverages a multi-scale attention mechanism. The model adopts a multi-stage feature extraction strategy: first, the preprocessed EEG signal is Fourier transformed to convert the time domain signal into a frequency domain representation; then, the deep neural network architecture is constructed. The first part of this neural network extracts harmonic features related to the stimulus frequency through the frequency domain convolutional layer; the second part uses spatial convolutional operations to capture topological correlations between electrodes in different brain regions; and in the third part, the pyramidal attention module is used to enhance the expression of key features through a multi-scale feature fusion and dynamic weight allocation mechanism. The final part of the neural network uses a time-domain convolutional network to mine the temporal dynamic properties of the signal. Ultimately, accurate classification is realized through the fully connected layer. In terms of training strategy, a three-stage progressive learning approach is adopted, where cross-participant common features are first learned, then fine-tuned for individuals, and finally the performance is evaluated on an independent test set.

Through rigorous comparative experiments and ablation analysis, the superiority of our proposed method over existing techniques is verified. We also confirm the key contribution of the Fourier transform and pyramid attention modules to the observed performance improvement. This study provides an important theoretical foundation and technical path for constructing a practical, robust and high-precision SSVEP-BCI system, and the proposed deep neural network method can also provide strong support for the research and application in related fields such as object detection and image processing.

REFERENCES

[1] Wolpaw J R. Brain-computer interfaces (BCIs) for communication and control[C]//Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility. 2007: 1-2.

[2] Jafri S R A, Hamid T, Mahmood R, et al. Wireless brain computer interface for smart home and medical system[J]. Wireless Personal Communications, 2019, 106: 2163-2177.

[3] Gao Q, Zhao X, Yu X, et al. Controlling of smart home system based on brain-computer interface[J]. Technology and Health Care, 2018, 26(5): 769-783.

[4] Abdulkader S N, Atia A, Mostafa M S M. Brain computer interfacing: Applications and challenges[J]. Egyptian Informatics Journal, 2015, 16(2): 213-230.

[5] Aznan N K N, Connolly J D, Al Moubayed N, et al. Using variable natural environment brain-computer interface stimuli for real-time humanoid robot navigation[C]//2019 international conference on robotics and automation (ICRA). IEEE, 2019: 4889-4895.

[6] Dai G, Zhou J, Huang J, et al. HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification[J]. Journal of neural engineering, 2020, 17(1): 016025.

[7] Amin S U, Alsulaiman M, Muhammad G, et al. Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification[J]. Ieee Access, 2019, 7: 18940-18950.

[8] Liu W, Guo C, Gao C. A cross-session motor imagery classification method based on Riemannian geometry and deep domain adaptation[J]. Expert Systems with Applications, 2024, 237: 121612.

[9] Li Y, Wang Y, Lei B, et al. SCDM: Unified representation learning for EEG-to-fNIRS cross-modal generation in MI-BCIs[J]. IEEE Transactions on Medical Imaging, 2025.

[10] Xie X, Chen L, Qin S, et al. Bidirectional feature pyramid attention-based temporal convolutional network model for motor imagery electroencephalogram classification[J]. Frontiers in Neurorobotics, 2024, 18: 1343249.

[11] Cui Y, Xie S, Xie X, et al. LDER: a classification framework based on ERP enhancement in RSVP task[J]. Journal of Neural Engineering, 2023, 20(3): 036029.

[12] Santamaria-Vazquez E, Martinez-Cagigal V, Vaquerizo-Villar F, et al. EEG-inception: a novel deep convolutional neural network for assistive ERP-based brain-computer interfaces[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2020, 28(12): 2773-2782.

[13] Leoni J, Strada S C, Tanelli M, et al. Automatic stimuli classification from ERP data for augmented communication via Brain–Computer Interfaces[J]. Expert Systems with Applications, 2021, 184: 115572.

[14] Deng Y, Ji Z, Wang Y, et al. OS-SSVEP: one-shot SSVEP classification[J]. Neural Networks, 2024, 180: 106734.

[15] Wong C M, Wang B, Wang Z, et al. Spatial filtering in SSVEP-based BCIs: Unified framework and new improvements[J]. IEEE Transactions on Biomedical Engineering, 2020, 67(11): 3057-3072.

[16] Yan W, Wu Y, Du C, et al. Cross-subject spatial filter transfer method for SSVEP-EEG feature recognition[J]. Journal of neural engineering, 2022, 19(3): 036008.

[17] Lin Z, Zhang C, Wu W, et al. Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs[J]. IEEE transactions on biomedical engineering, 2006, 53(12): 2610-2614.

[18] Chen X, Wang Y, Gao S, et al. Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface[J]. Journal of neural engineering, 2015, 12(4): 046008.

[19] Chen X, Wang Y, Nakanishi M, et al. High-speed spelling with a noninvasive brain–computer interface[J]. Proceedings of the national academy of sciences, 2015, 112(44): E6058-E6067.

[20] Zhang Y U, Zhou G, Jin J, et al. Frequency recognition in SSVEP-based BCI using multiset canonical correlation analysis[J]. International journal of neural systems, 2014, 24(04): 1450013.

[21] Nakanishi M, Wang Y, Chen X, et al. Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis[J]. IEEE Transactions on Biomedical Engineering, 2017, 65(1): 104-112.

[22] Wong C M, Wan F, Wang B, et al. Learning across multi-stimulus enhances target recognition methods in SSVEP-based BCIs[J]. Journal of neural engineering, 2020, 17(1): 016026.

[23] Zhang F, Panahi A, Gao G. FsaNet: Frequency self-attention for semantic segmentation[J]. IEEE Transactions on Image Processing, 2023, 32: 4757-4772.

[24] Gu J, Kwon H, Wang D, et al. Multi-scale high-resolution vision transformer for semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 12094-12103.

[25] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[26] Zhang R, Li J, Sun H, et al. Scan: Self-and-collaborative attention network for video person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(10): 4870-4882.

[27] Xiao S, Zhao Z, Zhang Z, et al. Query-biased self-attentive network for query-focused video summarization[J]. IEEE Transactions on Image Processing, 2020, 29: 5889-5899.

[28] O'shea K, Nash R. An introduction to convolutional neural networks[J]. arXiv preprint arXiv:1511.08458, 2015.

[29] Chen Z, He Z, Lu Z M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention[J]. IEEE transactions on image processing, 2024, 33: 1002-1015.

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

[31] Zhao D, Wang T, Tian Y, et al. Filter bank convolutional neural network for SSVEP classification[J]. IEEE Access, 2021, 9: 147129-147141.

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

[33] Liu J, Xie J, Zhang H, et al. Enhancing Steady-State Visual Evoked Potential Brain-Computer Interface Performance with Pyramid Squeeze Attention a Deep Learning Approach[C]//2024 6th International Conference on Electronic Engineering and Informatics (EEI). IEEE, 2024: 424-428.

[34] Waytowich N, Lawhern V J, Garcia J O, et al. Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials[J]. Journal of neural engineering, 2018, 15(6): 066031.

[35] Chen J, Zhang Y, Pan Y, et al. A transformer-based deep neural network model for SSVEP classification[J]. Neural Networks, 2023, 164: 521-534.

[36] Zhang X, Wei W, Qiu S, et al. Enhancing SSVEP-Based BCI Performance via Consensus Information Transfer Among Subjects[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024.

[37] Zhang H, Zu K, Lu J, et al. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network[C]//Proceedings of the asian conference on computer vision. 2022: 1161-1177.

[38] Jiang X, Cao B, Ou L, et al. iFuzzyTL: Interpretable Fuzzy Transfer Learning for Steady-State Visual Evoked Potentials Brain–Computer Interfaces System[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2025, DOI: 10.1109/TSMC.2025.3614244.

[39] Yang Y, Wang Z, Jia Z, et al. Dual-Branch Attention-based Frequency Domain Network for Cross-subject SSVEP-BCIs[J]. IEEE Journal of Biomedical and Health Informatics, 2025, DOI: 10.1109/JBHI.2025.3630249.

[40] Mao A, Mohri M, Zhong Y. Cross-entropy loss functions: Theoretical analysis and applications[C]//International conference on Machine learning. PMLR, 2023: 23803-23828.

[41] McFarland D J, Wolpaw J R. Brain-computer interfaces for communication and control[J]. Communications of the ACM, 2011, 54(5): 60-66.

[42] Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(Nov): 2579-2605.

[43] Wang Y, Chen X, Gao X, et al. A benchmark dataset for SSVEP-based brain–computer interfaces[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2016, 25(10): 1746-1752.

Xiao Wu obtained a bachelor's degree in Mathematics and Applied Mathematics from Dongchang College of Liaocheng University (located in Liaocheng City, China) in 2020, and later received a master's degree in Mathematics from the School of Mathematical Sciences of Liaocheng University in 2023. She is currently pursuing a doctoral degree in Control Science and Engineering at East China University of Science and Technology in Shanghai, China. Her main research interests include brain-computer interfaces, deep learning, and pattern recognition.

Ian Daly received the M.Eng. degree in computer science and the Ph.D. degree in cybernetics from the University of Reading, Reading, U.K., in 2006 and 2011, respectively. Between May 2011 and 2013 he was a Postdoctoral Researcher with the Laboratory of Brain-Computer Interfaces, Graz University of Technology, Graz, Austria. He is currently a Senior Lecturer with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K. His research interests focus on BCIs, nonlinear dynamics, machine learning, signal processing, and connectivity analysis in the EEG and fMRl (functional magnetic resonance imaging). He is also interested in the neurophysiological correlates of motor control and stimuli perception and how they differ between healthy participants and individuals with neurological and physiological impairments.

Andrew Ty Lau graduated from New York University. Currently, he works at Shanghai Lansheng Brain Hospital Investment Co., Ltd., mainly engaged in the research and development of brain-computer interfaces and clinical translation. His research focuses on brain-computer interfaces, target classification and recognition, deep learning algorithms, and neural signal processing. He is dedicated to promoting the application of cross-disciplinary technologies between artificial intelligence and brain science in clinical diagnosis, treatment, and rehabilitation.

Weijie Chen received the B.S. degree in Electronic Information Engineering from Anhui Polytechnic University, Wuhu, China, in 2020, and received the M.S. degree in Electronic Information with the School of Information and Electrical Engineering at Hunan University of Science and Technology in Xiangtan, China, in 2023. He is currently pursuing the Ph.D. in Control Science and Engineering at East China University of Science and Technology in Shanghai, China. His main research interests include brain-computer interfaces, neural networks, and robotics.

Chongfeng Wang received a bachelor's degree in Electronic Information Engineering from Weifang University (located in Weifang, China) in 2021, and subsequently obtained a master's degree in Electronic Information Engineering from the School of Optoelectronic Technology at Qilu University of Technology in 2024. He is currently pursuing a doctoral degree in Control Science and Engineering at East China University of Science and Technology in Shanghai, China. His main research interests include brain–computer interfaces, deep learning, and pattern recognition.

Andrzej Cichocki (Fellow, IEEE) received the M.Sc. (Hons.), Ph.D., and Dr.Sc. (Habilitation) degrees in electrical engineering from the Warsaw University of Technology, Warszawa, Poland, in 1972, 1975, and 1982, respectively.

He spent several years at the University Erlangen, Erlangen, Germany, as an Alexander-von-Humboldt Research Fellow and a Guest Professor. From 1995 to 2017, he was a Senior Team Leader and the Head of the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako, Japan. He is currently with the Systems Research Institute, Polish Academy of Science, and with RIKEN Advanced Intelligence Project, and also Tokyo University of Agriculture and Technology. Currently, his research focus on multiway blind source separation, tensor decomposition, tensor networks for big data mining, and brain-computer interface.

Jing Jin (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from the East China University of Science and Technology, Shanghai, China, in 2010. His Ph.D. advisors were Prof. Gert Pfurtscheller at Graz University of Technology from 2008 to 2010 and Prof. Xingyu Wang at East China University of Science and Technology from 2006 to 2008.

He is currently a Professor at East China University of Science and Technology (ECUST) and vice Dean of School of Mathematics at ECUST.

Prof. Jin currently serves as Associate Editor of Cognitive Neurodynamics, Journal of Neuroscience Methods and Frontiers in Neurorobotics, Action Editor of Neural Networks, Editor of Journal of Neural Engineering. His research interests include brain-computer interface, signal processing and pattern recognition.