A Novel Data Augmentation Approach Using Mask Encoding for Deep Learning-Based Asynchronous SSVEP-BCI

Wenlong Ding, *Graduate Student Member, IEEE*, Aiping Liu, *Member, IEEE*, Ling Guan, and Xun Chen, *Senior Member, IEEE*

***Abstract*— Deep learning (DL)-based methods have been successfully employed as asynchronous classification algorithms in the steady-state visual evoked potential (SSVEP)-based brain-computer interface (BCI) system. However, these methods often suffer from the limited amount of electroencephalography (EEG) data, leading to overfitting. This study proposes an effective data augmentation approach called EEG mask encoding (EEG-ME) to mitigate overfitting. EEG-ME forces models to learn more robust features by masking partial EEG data, leading to enhanced generalization capabilities of models. Three different network architectures, including an architecture integrating convolutional neural networks (CNN) with Transformer (CNN-Former), time domain-based CNN (tCNN), and a lightweight architecture (EEGNet) are utilized to validate the effectiveness of EEG-ME on publicly available benchmark and BETA datasets. The results demonstrate that EEG-ME significantly enhances the average classification accuracy of various DL-based methods with different data lengths of time windows on two public datasets. Specifically, CNN-Former, tCNN, and EEGNet achieve respective improvements of 3.18%, 1.42%, and 3.06% on the benchmark dataset as well as 11.09%, 3.12%, and 2.81% on the BETA dataset, with the 1-second time window as an example. The enhanced performance of SSVEP classification with EEG-ME promotes the implementation of the asynchronous SSVEP-BCI system, leading to improved robustness and flexibility in human-machine interaction.**

***Index Terms*— Asynchronous brain-computer interface, data augmentation, deep learning, electroencephalography mask encoding, steady-state visual evoked potential.**

Manuscript received 12 September 2023; revised 15 November 2023 and 25 January 2024; accepted 8 February 2024. Date of publication 19 February 2024; date of current version 22 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC3603600; in part by the National Natural Science Foundation of China under Grant 32271431, Grant 82272070, and Grant 82271516; and in part by the Fundamental Research Funds for the Central Universities under Grant KY2100000123. *(Corresponding author: Xun Chen.)*

Wenlong Ding, Aiping Liu, and Xun Chen are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: dingwenlong@mail.ustc.edu.cn; aipingl@ustc.edu.cn; xunchen@ustc.edu.cn).

Ling Guan is with the National Center for Neurological Disorders, Beijing Tiantan Hospital, Beijing 100070, China, and also with the Department of Medicine, The University of British Columbia, Vancouver, BC V6H 3N1, Canada (e-mail: lguanm@hotmail.com).

Digital Object Identifier 10.1109/TNSRE.2024.3366930

I. Introduction

BRAIN-COMPUTER interface (BCI) allows individuals to communicate silently with the outside world without the need for sound or movement [1]. It establishes a direct communication channel between the brain and external device based on the individual's neural activity [2]. Electroencephalography (EEG) is a commonly used neural recording technique with the advantages of convenience and safety for BCI applications [3], [4], [5], and Steady-state visual evoked potential (SSVEP) is widely adopted as one of the most promising EEG paradigms for BCI [6]. The SSVEP signal demonstrates superior stability, and its frequency domain features are distinctly observable. When employing flickering visual stimuli at various frequencies, it's possible to generate SSVEP-EEG signals that contain frequency components corresponding to the targeted stimuli. By decoding SSVEP-EEG signals, it becomes feasible to identify the specific target that the user is actively focusing on and facilitates effective communication for BCI [7].

BCI can be classified into two modes: synchronous and asynchronous [6], [7], [8], [9], [10]. The synchronous BCI system is characterized by the use of predefined time windows with a specific cue or trigger that indicates the onset of mental activity. As a result, synchronous systems limit user activities to adhere to the designated time sequence [6]. In contrast, the asynchronous BCI system offers greater flexibility, enabling users to issue commands according to their intentions at any time. This system does not require any predefined time windows and can continuously decode the user's intentions [9].

Various methods have been proposed to assist in decoding SSVEP-EEG. Traditional methods include canonical correlation analysis (CCA)-derived methods [10], [11], [12], [13], [14], task-related component analysis (TRCA)-derived methods [15], [16], and task-discriminant component analysis (TDCA) [17]. Deep learning (DL) [18] methods include deep neural network (DNN) [19], SSVEPformer [20], and generalized zero-shot learning (GZSL) [21]. During the training phase of these methods, specific-position time windows in the stimulus trials are selected as training samples, such as $[0.14, 0.14 + d_s]$ s, which starts at 0.14 s after stimulus onset, and d_s s denotes the data length of time windows.

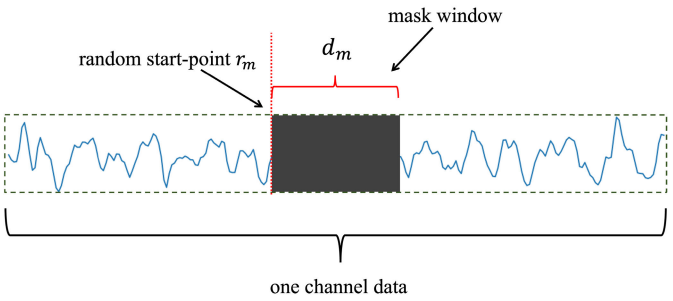

Fig. 1. EEG-ME for one electrode channel data of a sample.

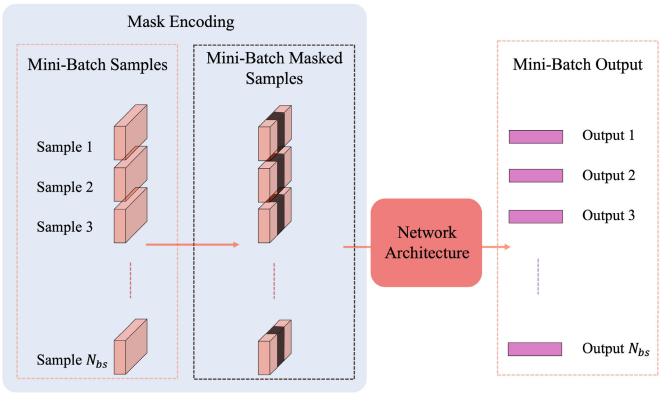

Fig. 2. The application of EEG-ME during the training phase.

The same specific-position time window is required in the test phase. These methods, which exhibit impressive classification (decoding) performance within predefined time windows synchronized with the flicker of stimulus targets, can be termed synchronous SSVEP classification algorithms. However, they exhibit poor performance outside of predefined time windows [7]. Thus implementing asynchronous SSVEP-BCI system using synchronous SSVEP classification algorithms is challenging.

Some methods exhibit stable classification performance at any time that can be termed asynchronous SSVEP classification algorithm. Traditional methods, such as CCA [23] (and its filter bank version [24]), multivariate synchronization index (MSI) [25], and Ramanujan periodicity transforms (RPT) [26], can implement asynchronous SSVEP classification, but a longer time window is required. In recent years, some DL-based methods have been proposed for asynchronous SSVEP classification, such as Fast Fourier Transform (FFT)-based convolutional neural networks (CNN) [27], [28], time domain-based CNN (tCNN) [7], and EEGNet [29]. They can achieve superior classification performance using shorter time windows in comparison to traditional methods [7]. To build a fast and flexible asynchronous SSVEP-BCI system in the future, we focus on the DL-based asynchronous SSVEP classification algorithm in this study.

However, the classification performance of DL-based methods is constrained by the limited amount of EEG data [30], [35], [37]. For public datasets such as benchmark [31] and BETA [32], each stimulus target contains only six or four trials, respectively. DL-based methods tend to overfit on these datasets, leading to a decrease in classification performance. Data augmentation is a promising strategy that can effectively prevent overfitting [33]. Generative adversarial networks (GAN) and variational auto-encoders (VAE) have been used to generate additional EEG samples to enhance the SSVEP classification performance [35]. However, GAN and VAE would introduce additional training costs. SpecAugment, which is created for data augmentation in speech recognition [36], has been attempted in the SSVEP classification [37]. It first transfers the EEG signals into the spectrogram images by Short-Time Fourier Transform (STFT) and then applies frequency masking and time masking to it. However, some useful information in the time domain may be missed by the STFT, and the complex operation of SpecAugment would introduce additional time costs [37].

In this study, we propose a novel data augmentation approach called **EEG** **M**ask **E**ncoding (EEG-ME) to enhance the performance of DL-based SSVEP classification algorithms. A randomly positioned continuous segment is chosen as the mask window, and the value of the EEG data in the mask window is set to 0 (Fig. 1). EEG-ME makes it challenging for the model (network architecture) to extract features, forcing the model to learn more robust features from the data. This enables the model to better comprehend the underlying patterns in the EEG data rather than relying on memorization. Consequently, EEG-ME effectively mitigates overfitting and enhances the model's generalization capabilities. The proposed EEG-ME resembles Dropout [34] but differs in that it operates on a continuous segment region of the input data. EEG-ME can be applied directly to the EEG samples during the training phase (Fig. 2) and does not require any additional parameter learning. The higher the mask ratio (the proportion of mask window in the whole sample), the better the suppression of overfitting. However, as the mask ratio increases, so does the loss of information. Therefore, it is critical to determine the best mask ratio of the EEG-ME that maintains the trade-off between the reduction of overfitting and the loss of information.

To evaluate the effectiveness of the proposed EEG-ME approach, we combine EEG-ME with different DL-based methods and conduct experiments on the benchmark and BETA datasets respectively. Two state-of-the-art (SOTA) network architectures for asynchronous SSVEP classification, namely tCNN [7] and EEGNet [29] are employed in this study. tCNN is a standard CNN architecture without pooling layers and EEGNet is a lightweight architecture. Given the success of Transformer [38] in learning task-related features in natural language process (NLP) [39], computer vision (CV) [40], and EEG fields [20], [41], [42], we propose a novel network architecture that integrates CNN with Transformer (CNN-Former) for asynchronous SSVEP classification to further validate the benefits of EEG-ME.

The main contributions of this paper are summarized as follows.

1) We propose EEG-ME, a simple yet effective data augmentation approach for DL-based asynchronous SSVEP classification tasks. EEG-ME does not require additional parameter learning and can be easily implemented.

2) We validate the effectiveness of EEG-ME on different public datasets using various DL-based methods with different mask ratios and data lengths of time windows.
3) We conduct a comprehensive analysis of the factors that impact the effectiveness of EEG-ME and delve into the underlying theories that contribute to its effectiveness.
4) We employ frame-by-frame analysis to illustrate the distinction between asynchronous and synchronous algorithms as well as highlight the significance of asynchronous algorithms in the asynchronous BCI system.

The structure of this paper is outlined as follows. Section I presents a brief introduction to SSVEP algorithms and data augmentation approach. Section II introduces the proposed EEG-ME approach and CNN-Former architecture. The experiments and results are elaborated in Section III. Section IV presents the discussion on the effectiveness of the proposed EEG-ME approach and analyzes the significance of asynchronous algorithms in the asynchronous BCI system. Finally, Section V concludes this paper.

II. Methods

A. Datasets

Two widely-used and reliable public datasets, i.e., benchmark [31] and BETA [32] are used to evaluate the effectiveness of the proposed EEG-ME approach.

1) Benchmark Dataset: Thirty-five subjects participated in this experiment. Forty targets are coded using a joint frequency and phase modulation (JFPM) approach. Specifically, the stimulation frequency range is 8 to 15.8 Hz with an interval of 0.2 Hz, and the phase range is 0 to 1.5 π with an interval of 0.5 π. The EEG data of a subject consists of six blocks. Each block contains 40 trials in random order. Each trial begins with a 0.5 s target-cue stage. Next comes the stimulus stage, all stimulus targets start flashing simultaneously on the screen for 5 s. Followed by the rest stage, the screen goes blank for 0.5 s before the next trial begins. EEG data is recorded by the 64-channel device and downsampled to 250 Hz. The average visual delay across all subjects is 0.14 s. For more information, please refer to [31].

2) BETA Dataset: Seventy subjects participated in this experiment. The stimulus paradigm design of the BETA dataset shares similarities with the benchmark dataset, but it also has certain important differences. The BETA dataset is developed for real-world applications, which consists of data collected outside the laboratory setting of the electromagnetic shielding room. The BETA dataset consists of four blocks. The stimulus stage of each trial lasted 2 s for the first 15 subjects and 3 s for the last 55 subjects. Therefore, the useful EEG data in the BETA dataset is smaller than that in the benchmark dataset. The average visual delay across all subjects is 0.13 s. For more information, please refer to [32].

Considering the visual delay, the temporal range of useful data for each trial is $[0.5 + t_d, 0.5 + t_d + L]$ s, which starts $0.5+t_d$ s after the onset time of a trial, 0.5 s denotes the lasted time of cue stage, t_d denotes the average visual delay, and L s denotes the lasted time of the stimulus stage, as shown in Fig. 3. For the benchmark dataset, $t_d = 0.14$ s and $L = 5$ s.

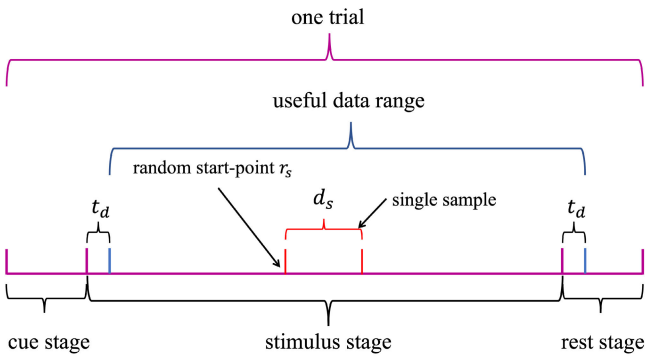

Fig. 3. The random selection process of a single sample in a trial.

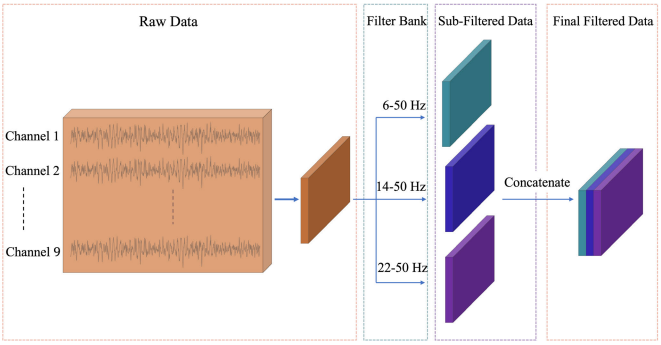

Fig. 4. Filter bank applied to the raw EEG data.

For the BETA dataset, $t_d = 0.13$ s, and $L = 2$ s for the first 15 subjects as well as $L = 3$ s for the last 55 subjects.

B. Pre-Processing

This study utilizes EEG data from nine electrode channels located in the occipital region, including Pz, PO5, PO3, POz, PO4, PO6, O1, Oz, and O2 [31], [32]. To enhance the efficiency of the network architecture for learning task-related features, a filter bank is employed to filter the data. The filter bank comprises several sub-filters with different bandpass ranges that can utilize the fundamental and harmonic information of SSVEP-EEG data more effectively [7], [19], [24]. In both the benchmark and BETA datasets, the fundamental range is 8-15.8 Hz, the second harmonic range is 16-31.6 Hz, and the third harmonic range is 24-47.4 Hz. Harmonic information above 50 Hz is not used [7] since the signal-to-noise ratio (SNR) is relatively low above 50 Hz for the benchmark and BETA datasets used in this study. Following method M_3 in [24], the three sub-filters with different bandpass ranges are designed as 6-50 Hz, 14-50 Hz, and 22-50 Hz respectively. Sixth-order Butterworth filters are employed as the sub-filters in this study. The filtered data from the three sub-filters are concatenated to obtain the final filtered data as in [19]. Fig. 4 shows the filter bank applied to the raw EEG data.

C. Sample Selection Process

To create an asynchronous SSVEP classification algorithm, random-position time windows in the useful data range are selected as training samples [7], with a shape of $3@9\times d_s$

(three sub-filters@nine electrode channels × data length of time windows). The temporal range of time windows is $[t_d + r_s, t_d + r_s + d_s]$ s, which starts $t_d + r_s$ s after the onset time of the stimulus and r_s denotes a random number between $[0, L - d_s]$. The random selection process of a single sample is shown in Fig. 3.

D. The Proposed EEG-ME Approach

During the training phase, a mask window is applied to the selected sample. The value of the data in the mask window is set to 0. Fixing the position of the mask window hinders the model's ability to learn how to handle data in that specific position. Consequently, the partial data within the specific position of test samples does not contribute significantly to the decoding process of the model. To address this issue, randomizing the position of the mask window is necessary.

Fig. 1 illustrates the EEG-ME process for a single sample, focusing on one electrode channel. The mask window has a temporal range of $[r_m, r_m + d_m]$ s, where r_m represents the start point of mask windows, $d_m = d_s \times ratio_m$ s denotes the data length of mask windows, and $ratio_m$ denotes the mask ratio of samples. r_m is a random value within the range $[0, d_s \times (1 - ratio_m)]$ s. In terms of program implementation, the operation unit is the frame. Thus the range should be multiplied by the data sampling rate, resulting in a specific value range of $[0, 250 \times d_s \times (1 - ratio_m)]$. It is important to note that all nine electrode channels of a sample share the same mask window.

Network weights are updated every other mini-batch, with each mini-batch containing N_{bs} (N_{bs} = mini-batch size) samples. r_m can employ two variation strategies: (1) Random variation of r_m with individual samples, and (2) Setting r_m to a consistent value for the N_{bs} samples in a mini-batch, while randomly varying it across different mini-batches. Both strategies result in equivalent generalization capabilities of models (Fig. 8) but the latter simplifies computations. Hence, this study adopts strategy (2). The process of the proposed EEG-ME approach applied to the EEG samples during the training phase is shown in Fig. 2. It is worth noting that during the training phase, EEG-ME serves as an online augmentation technique applied to EEG samples.

E. State-of-the-Art DL-Based Asynchronous Algorithms

tCNN [7] and EEGNet [29] are employed to validate the effectiveness of EEG-ME. Table I illustrates the architecture of tCNN. To adapt to the classification task of a great number of categories, the output dimension in the convolutional layer is larger than that in our previous work [7]. l_1 and l_2 represent the temporal length of the front layer. Table II illustrates the architecture of EEGNet. Step sizes in EEGNet are set to 1.

TABLE I
THE ARCHITECTURE OF TCNN

Layer type	Output dimension	Kernel size	Step size	Options
Input				
Conv2D	8	9×1	1	mode=valid
BN				
Activation				ELU
Droupout				ratio=0.5
Conv2D	128	1×l_1	5	mode=same
BN				
Activation				ELU
Droupout				ratio=0.5
Conv2D	128	1×5	1	mode=valid
BN				
Activation				ELU
Droupout				ratio=0.5
Conv2D	128	1×l_2	1	mode=valid
BN				
Activation				ELU
Flatten				
Droupout				ratio=0.5
Dense	40			
Activation				softmax

TABLE II
THE ARCHITECTURE OF EEGNET

Layer type	Output dimension	Kernel size	Options
Input			
Conv2D	96	1×l_1	mode=same
BN			
DepthwiseConv2D	96	9×1	mode=valid
BN			
Activation			ELU
AveragePool2D		1×4	
Droupout			ratio=0.5
SeparableConv2D	96	1×16	mode=same
BN			
Activation			ELU
AveragePool2D		1×8	
Droupout			ratio=0.5
Flatten			
Dense	40		
Activation			softmax

F. The Proposed CNN-Former Architecture

Transformer has been successfully applied in EEG research [20], [41], [42], with specific focus and detailed discussion in [42]. To enhance the reliability of the validation results of EEG-ME, we propose the CNN-Former integrating CNN with Transformer, whose architecture is very different from tCNN and EEGNet. The code of CNN-Former with EEG-ME is available for reproducibility at https://github.com/DingWenl/CNN-FormerWithEEG-ME. Fig. 5(a) shows the architecture of the CNN-Former. The CNN module captures temporal and spatial features, while the Transformer module learns global temporal dependencies [42]. A multi-scale block is added to the CNN module to learn multi-scale information [44]. A fully connected layer with softmax is used to obtain the scores for the forty categories, and the category with the highest score is identified as the predicted category.

1) CNN Module: The input data is first filtered by a convolution layer with 9 × 1 convolution kernels, which enables each output unit to contain the spatial information of the nine electrodes. Then, the feature maps pass through a multi-scale block, which is inspired by the res2net [44], as shown in Fig. 5(b). Different convolution kernels are designed to learn

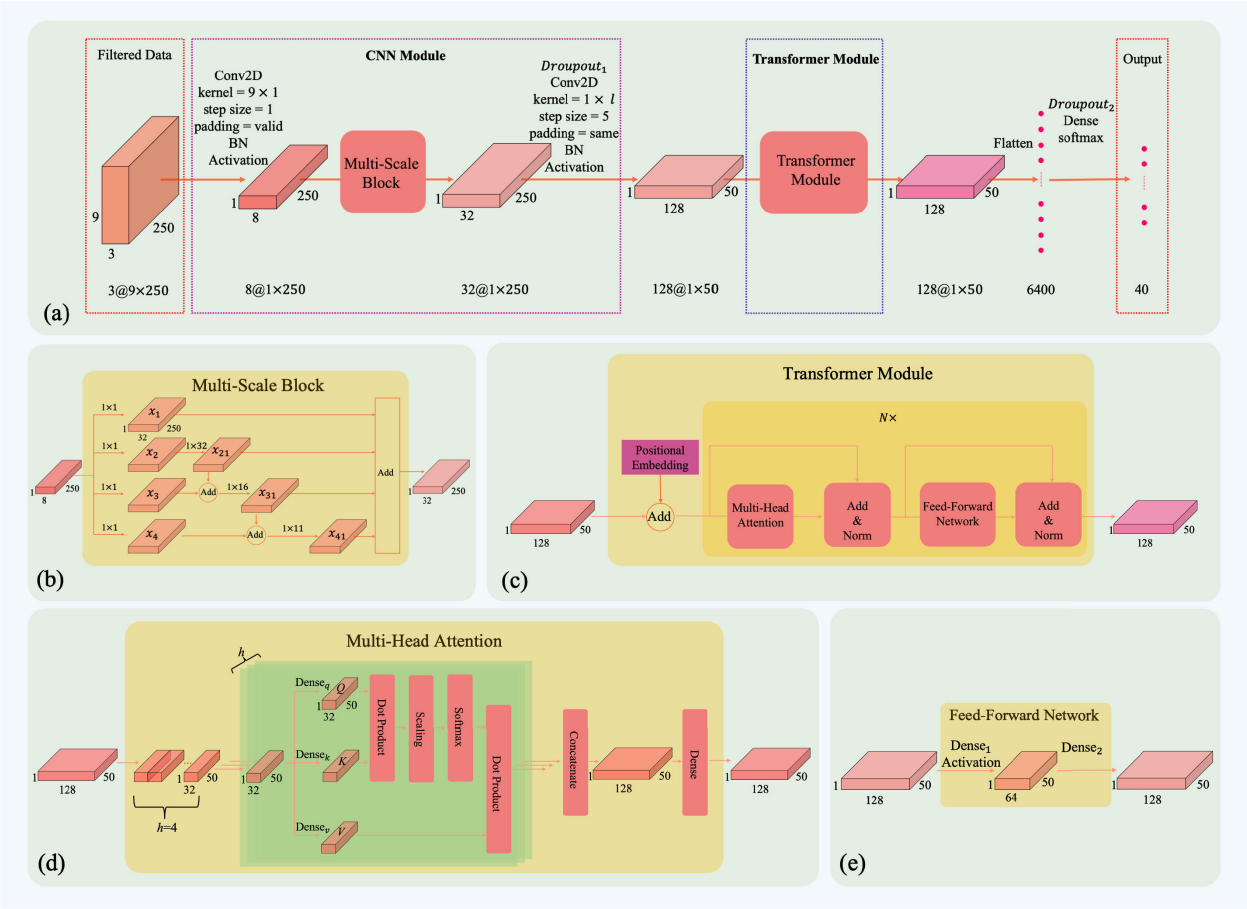

Fig. 5. The illustration of the proposed CNN-Former using the example input data of a 1-second time window. (a) CNN-Former architecture, *Conv2D* denotes two-dimensional convolution, *BN* denotes batch normalization, *Dense* denotes the fully connected layer, *Activation* = ELU [43], $Dropout_1 = 0.5$, $Dropout_2 = 0.95$, *padding* = *same/valid* denotes padding is used or not used; (b) Multi-scale block; (c) Transformer module; (d) Multi-head attention mechanism, *h* denotes the number of heads; (e) Feed-forward network.

the information of different temporal scales. Next, a convolution layer with $1 \times l$ convolution kernels is utilized, where l denotes the data length of the input. The large kernel size ensures that each output unit contains sufficient temporal information.

2) Transformer Module: As shown in Fig. 5(c), the Transformer module contains a positional encoding and an encoder with a stack of N identical layers [38]. Each layer has two sub-layers: a multi-head attention mechanism (Fig. 5(d)) and a fully connected feed-forward network (Fig. 5(e)). A residual connection [45] is employed around each of the two sub-layers, followed by layer normalization [46]. Dropout is applied to the output of each sub-layer before being added to the sub-layer input and normalized. Additionally, dropout is applied to the sums of inputs and positional encodings. For more detailed information, please refer to [38]. In this study, the Transformer module is configured with the following parameters: $N = 2$ and dropout ratio $= 0.1$.

G. Performance Evaluation

Accuracy, which is defined as the ratio of the number of correct samples to the number of total samples, is used to evaluate the classification performance of DL-based methods. The accuracy P is expressed as:

$$P = n_1/n, \tag{1}$$

where n_1 denotes the number of correct samples, and n denotes the number of total samples.

Information transfer rate (ITR) is an important index in the BCI system [17]. ITR considers the trade-off between accuracy and data length. ITR (bits/min) is defined as:

$$ITR = 60 \times [log_2 M + P log_2 P + (1-P) log_2 \frac{1-P}{M-1}]/d_s, \tag{2}$$

where M denotes the number of stimulus targets, P denotes the accuracy, and d_s s denotes the data length of test samples.

The paired t-test is conducted to examine any significant differences in average classification accuracy or ITR between each pair of methods under each condition. In the event of a significant main effect ($p < 0.05$), post hoc t-test comparisons are subsequently conducted with the application of the Bonferroni correction.

III. Experiments and Results

A. Experimental Setup

1) Comparative Experiments: The experiments encompass five parts: (1) Comparative experiments that involve various EEG-ME-based DL methods with different mask ratios and data lengths of time windows, (2) Comparative experiments assessing the classification performance of the enhanced

TABLE III
AVERAGE CLASSIFICATION ACCURACY AND ITR (MEAN ± SEM) OF DL-BASED METHODS WITH DIFFERENT MASK RATIOS AND DATA LENGTHS OF TIME WINDOWS ON BENCHMARK DATASET

		0.8 s		1 s		1.2 s	
Methods	**Mask Ratio**	**Accuracy (%)**	**ITR (bits/min)**	**Accuracy (%)**	**ITR (bits/min)**	**Accuracy (%)**	**ITR (bits/min)**
CNN-Former	0	74.11 ± 2.68	241.27 ± 12.73	82.65 ± 2.35	229.51 ± 9.80	88.10 ± 2.06	212.45 ± 7.59
	0.1	76.17 ± 2.53***	251.46 ± 12.33***	84.29 ± 2.22***	236.75 ± 9.46***	89.71 ± 1.83***	218.63 ± 6.95***
	0.2	77.61 ± 2.39***	258.57 ± 11.84***	85.62 ± 2.06***	242.55 ± 8.94***	90.66 ± 1.68***	222.36 ± 6.53***
	0.3	**77.76 ± 2.42*** **	**259.46 ± 11.88*** **	**85.83 ± 2.07*** **	**243.59 ± 8.97*** **	**90.69 ± 1.68*** **	**222.52 ± 6.53*** **
	0.4	76.57 ± 2.52***	253.39 ± 12.09***	84.95 ± 2.18***	239.66 ± 9.23***	90.19 ± 1.80***	220.64 ± 6.86***
	0.5	71.88 ± 3.07*	231.55 ± 13.26**	81.90 ± 2.66	226.99 ± 10.46	87.81 ± 2.17	211.48 ± 7.78
tCNN	0	73.52 ± 2.53	237.54 ± 12.07	81.47 ± 2.31	223.86 ± 9.57	86.84 ± 2.06	207.18 ± 7.56
	0.1	**74.28 ± 2.47*** **	**241.18 ± 11.87*** **	82.68 ± 2.26***	229.31 ± 9.49***	87.94 ± 1.98***	211.57 ± 7.40***
	0.2	73.82 ± 2.54	239.11 ± 12.12	**82.89 ± 2.23*** **	**230.13 ± 9.35*** **	**88.57 ± 1.91*** **	**213.99 ± 7.17*** **
	0.3	72.11 ± 2.59***	230.45 ± 12.04***	81.98 ± 2.32	226.24 ± 9.62*	88.29 ± 1.92***	212.83 ± 7.19***
	0.4	68.20 ± 2.72***	211.59 ± 12.09***	78.98 ± 2.52***	213.37 ± 10.01***	86.51 ± 2.15	206.10 ± 7.77
	0.5	62.59 ± 2.81***	185.58 ± 11.79***	73.90 ± 2.90***	192.89 ± 10.65***	82.22 ± 2.57***	190.17 ± 8.59***
EEGNet	0	64.38 ± 2.73	193.57 ± 11.97	72.78 ± 2.84	188.21 ± 10.71	78.76 ± 2.76	177.92 ± 9.13
	0.1	65.58 ± 2.70***	199.07 ± 11.94***	74.42 ± 2.78***	194.78 ± 10.64***	80.44 ± 2.66***	183.86 ± 8.95***
	0.2	**67.29 ± 2.68*** **	**207.11 ± 12.07*** **	**75.84 ± 2.63*** **	**200.18 ± 10.21*** **	**81.64 ± 2.55*** **	**188.05 ± 8.71*** **
	0.3	67.27 ± 2.67***	206.96 ± 11.92***	75.79 ± 2.67***	200.10 ± 10.29***	81.60 ± 2.57***	187.89 ± 8.66***
	0.4	65.52 ± 2.71**	198.73 ± 11.86**	74.30 ± 2.74***	194.07 ± 10.41***	80.13 ± 2.64***	182.56 ± 8.79***
	0.5	57.42 ± 2.54***	161.52 ± 10.36***	66.78 ± 2.73***	163.88 ± 9.65***	74.15 ± 2.76***	161.24 ± 8.65***

Note: The asterisks denote significant differences between each mask ratio and zero mask ratio by paired t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

CNN-Former in contrast to the most commonly utilized traditional asynchronous classification methods, (3) Experiments comparing EEG-ME with two commonly employed data augmentation techniques for EEG decoding, (4) Experiments involving CNN-Former trained with and without EEG-ME, employing different numbers of training blocks, and (5) Comparative experiments appraising two variation strategies of EEG-ME.

2) Model Training Strategy: In this study, Categorical Cross-entropy and Adam are selected as the loss function and optimization algorithm. Mini-batch size is set to 256, and the model is trained for 4000 mini-batches (iterations) on the benchmark dataset. For the BETA dataset, the number of mini-batches is set to 2000 due to its smaller original sample size. Finally, the trained model is recorded with the minimum loss of the training set.

3) Leave-One-Out Cross-Validation: Leave-one-out cross-validation is employed for the evaluation of SSVEP classification algorithms. The EEG data of a subject comprises N_b blocks, where one block is designated as the test set while N_{tb} randomly chosen blocks serve as the training set. This procedure is repeated N_b times, and the average of the resulting N_b classification results determines the subject's classification accuracy. For the benchmark dataset, N_b is set as 6, whereas for the BETA dataset, N_b is set as 4. In the experiment (4), N_{tb} varies as 1, 2, 3, 4, and 5 for the benchmark dataset, as well as 1, 2, and 3 for the BETA dataset. In other experiments, N_{tb} is fixed as 5 for the benchmark dataset and 3 for the BETA dataset. During the classification process, 5000 test samples are randomly selected from the relevant data range of the test set to evaluate the trained model.

B. Results

In this subsection, we use the following notations: P_{base} represents the average classification accuracy of DL-based methods without EEG-ME, while P_{best} represents the maximum average classification accuracy of DL-based methods with EEG-ME, and the corresponding mask ratio is defined as $ratio_b$. Furthermore, we define ΔP as the accuracy improvement from P_{base} to P_{best}, which is given by $\Delta P = P_{best} - P_{base}$. The accuracy improvement can reveal the effectiveness of EEG-ME. In addition, "w/ EEG-ME" and "w/o EEG-ME" respectively indicate CNN-Former with and without EEG-ME.

Table III and Table IV illustrate the average classification accuracy and ITR of CNN-Former, tCNN, and EEGNet across all subjects on the benchmark and BETA datasets, respectively, with different mask ratios and data lengths of the time windows. The mask ratio ranges from 0 to 0.5 with an interval of 0.1, where mask ratio = 0 indicates that EEG-ME is not applied to the DL-based methods. Considering the trade-off between accuracy and ITR, the data length of 1 s is commonly used in DL-based asynchronous SSVEP-BCI [28], [29]. Thus the primary analysis range of data lengths in this study from 0.8 to 1.2 s with an interval of 0.2 s. The paired t-test with Bonferroni correction reveals that EEG-ME can significantly improve the classification performance of DL-based methods with appropriate mask ratios ($p < 0.001$). As the mask ratio increases, the average classification accuracy and ITR first increase and then decrease. The maximum ITR for CNN-Former, tCNN, and EEGNet are 259.26 ± 11.88 bits/min, 241.18 ± 11.87 bits/min, and 207.11 ± 12.07 bits/min in the benchmark dataset, as well as 176.74 ± 11.98 bits/min, 165.86 ± 10.90 bits/min, and 146.83 ± 9.46 bits/min in the BETA dataset.

Additional experiments are conducted using CNN-Former to validate the effectiveness of EEG-ME. The data length range is extended from 0.2 to 1.2 s at 0.2 s intervals. The $ratio_b$ values for EEG-ME with CNN-Former are 0, 0, and 0.2 for 0.2 s, 0.4 s, and 0.6 s, respectively, on the benchmark dataset. Similarly, on the BETA dataset, the $ratio_b$ values are 0, 0.1, and 0.2 for the same data lengths. To better showcase the classification performance of CNN-Former improved by EEG-ME, its P_{best} and P_{base} are compared with the most commonly used SOTA traditional asynchronous algorithms, i.e., CCA [23] and FBCCA [24]. Fig. 6 illustrates the average classification accuracy of CNN-Former with and without EEG-ME,

TABLE IV
AVERAGE CLASSIFICATION ACCURACY AND ITR (MEAN ± SEM) OF DL-BASED METHODS WITH DIFFERENT MASK RATIOS AND DATA LENGTHS OF TIME WINDOWS ON BETA DATASET

		0.8 s		1 s		1.2 s	
Methods	**Mask Ratio**	**Accuracy (%)**	**ITR (bits/min)**	**Accuracy (%)**	**ITR (bits/min)**	**Accuracy (%)**	**ITR (bits/min)**
CNN-Former	0	50.76 ± 2.74	141.15 ± 10.64	53.97 ± 3.00	125.49 ± 9.70	55.46 ± 3.30	111.03 ± 9.07
	0.1	55.65 ± 2.79***	161.75 ± 11.20***	60.74 ± 3.01***	149.12 ± 10.14***	62.82 ± 3.11***	131.53 ± 9.02***
	0.2	58.56 ± 2.89***	175.29 ± 11.85***	64.90 ± 3.00***	164.55 ± 10.44***	68.77 ± 3.09***	150.40 ± 9.15***
	0.3	**58.80 ± 2.92*** **	**176.74 ± 11.98*** **	**65.06 ± 3.06*** **	**165.60 ± 10.59*** **	**69.44 ± 3.09*** **	**152.69 ± 9.20*** **
	0.4	55.62 ± 3.07***	164.54 ± 12.08***	63.69 ± 3.14***	161.17 ± 10.52***	68.41 ± 3.07***	149.11 ± 9.14***
	0.5	48.00 ± 2.97*	132.81 ± 10.99*	57.91 ± 3.17***	140.42 ± 10.24***	64.23 ± 3.14***	135.95 ± 8.90***
tCNN	0	55.17 ± 2.57	157.68 ± 10.49	60.35 ± 2.74	145.53 ± 9.47	63.49 ± 2.89	131.96 ± 8.55
	0.1	56.59 ± 2.61***	164.06 ± 10.73***	62.68 ± 2.73***	154.04 ± 9.57***	66.51 ± 2.85***	141.30 ± 8.59***
	0.2	**56.89 ± 2.66*** **	**165.86 ± 10.90*** **	**63.47 ± 2.78*** **	**157.31 ± 9.68*** **	**68.12 ± 2.83*** **	**146.46 ± 8.63*** **
	0.3	55.61 ± 2.72	160.93 ± 10.92	62.57 ± 2.81***	154.27 ± 9.69***	66.83 ± 2.81***	142.10 ± 8.44***
	0.4	49.57 ± 2.67***	135.85 ± 10.12***	58.59 ± 2.81**	139.79 ± 9.40**	63.77 ± 2.88	132.73 ± 8.35
	0.5	41.47 ± 2.59***	105.07 ± 9.11***	50.86 ± 2.80***	113.77 ± 8.57***	57.30 ± 2.87***	113.20 ± 7.84***
EEGNet	0	51.35 ± 2.28	139.55 ± 9.09	56.35 ± 2.52	129.82 ± 8.47	59.87 ± 2.70	119.65 ± 7.87
	0.1	**53.04 ± 2.33*** **	**146.83 ± 9.46*** **	58.53 ± 2.49***	137.25 ± 8.57***	62.23 ± 2.66***	126.58 ± 7.92***
	0.2	52.92 ± 2.31***	146.14 ± 9.31***	**59.16 ± 2.51*** **	**139.60 ± 8.65*** **	**62.65 ± 2.57*** **	**127.30 ± 7.64*** **
	0.3	51.52 ± 2.29	140.30 ± 9.06	56.94 ± 2.46	131.48 ± 8.36	60.80 ± 2.53*	121.39 ± 7.40
	0.4	46.89 ± 2.18***	121.31 ± 8.30***	52.31 ± 2.34***	115.12 ± 7.52***	55.81 ± 2.51***	106.61 ± 7.02***
	0.5	39.61 ± 1.94***	93.29 ± 6.77***	44.73 ± 2.17***	90.54 ± 6.52***	48.59 ± 2.25***	85.61 ± 5.87***

Note: The asterisks denote significant differences between each mask ratio and zero mask ratio by paired t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

TABLE V
AVERAGE CLASSIFICATION ACCURACY (MEAN ± SEM) OF CNN-FORMER WITH EEG-ME AND OTHER TWO DATA AUGMENTATION METHODS ON BENCHMARK AND BETA DATASETS

	Benchmark	**BETA**
Original	82.65 ± 2.35	53.97 ± 3.00
Gaussian Noise	82.32 ± 2.41***	55.90 ± 2.91***
Flipping	80.44 ± 2.58***	56.02 ± 2.83***
EEG-ME (ours)	**85.83 ± 2.07*** **	**65.06 ± 3.06*** **

Note: The black and red asterisks respectively denote significant differences between the original method and data augmentation methods, as well as between EEG-ME and the other two data augmentation methods by paired t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

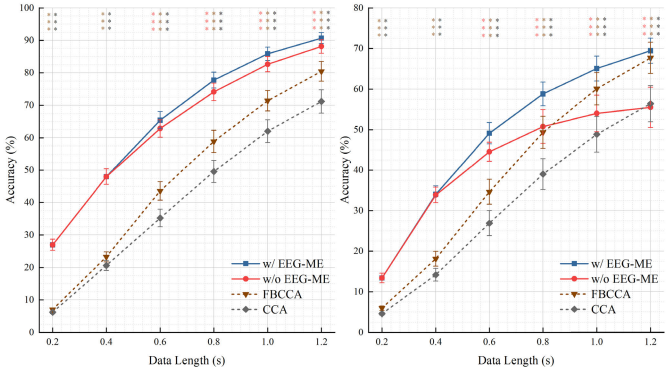

Fig. 6. Average classification accuracy of CNN-Former with and without EEG-ME, CCA, and FBCCA on (a) benchmark and (b) BETA datasets with different data lengths of time windows. The error bars denote SEM (standard error of the mean). Red, brown, and black asterisks respectively denote significant differences between "w/ EEG-ME" and "w/o EEG-ME", "w/ EEG-ME" and FBCCA, as well as "w/ EEG-ME" and CCA by paired t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

CCA, and FBCCA across all subjects on benchmark and BETA datasets with different time window lengths. Results suggest the effectiveness of EEG-ME applied to CNN-Former is more prominent after 0.4 s. In Table III, Table IV, and Fig. 6, the paired t-test with Bonferroni correction reveals that CNN-Former with EEG-ME significantly outperforms the other methods ($p < 0.001$), demonstrating its SOTA performance.

To establish the superiority of EEG-ME, we compare it with two commonly used EEG data augmentation techniques: Gaussian noise addition and flipping [33], [47]. Gaussian noise with an SNR of 5, similar to [47], is added to the EEG data. Flipping, which reverses the time dimension order, is also employed. Training samples have a 50% probability of flipping. SpecAugment is not used as it alters the input data shape, which is not suitable for the architectures used in this study. Table V illustrates the average classification accuracy of CNN-Former with EEG-ME and other two data augmentation methods on the benchmark and BETA datasets, using a data length of 1 s. The paired t-test with Bonferroni correction reveals that EEG-ME exhibits significantly superior performance compared to the other data augmentation techniques ($p < 0.001$). A noteworthy observation is that the other two methods demonstrate poor performance when applied to the benchmark dataset. The BETA dataset, which is collected in an open-world setting with elevated noise levels, exhibits significant disparities between the training and test sets that originate from the same subject. Introducing new samples with the other two methods may enhance the model's generalization capabilities on the test set. In contrast, the benchmark dataset boasts higher data quality and less noise, resulting in fewer disparities between the training and test sets derived from the same subject. Consequently, the introduced new samples may be redundant or lead to overfitting, ultimately undermining the model's performance.

To assess the impact of the number of original samples on the effectiveness of EEG-ME, we train the CNN-Former with and without EEG-ME using different numbers of blocks. Fig. 7 illustrates the average classification accuracy of "w/ EEG-ME" (mask ratios = 0.3) and "w/o EEG-ME" across all subjects on the benchmark and BETA datasets with various training blocks, using a data length of 1 s. The paired t-test

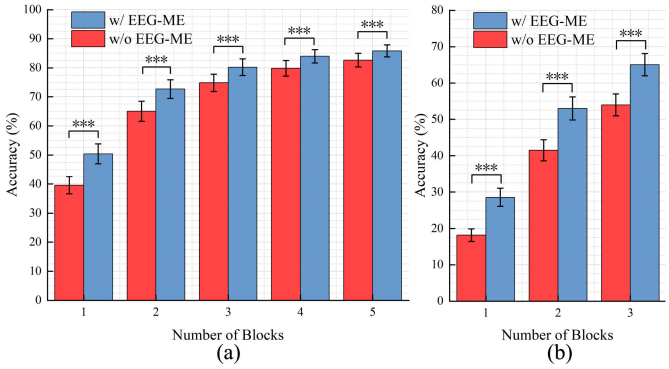

Fig. 7. Average classification accuracy of CNN-Former with and without EEG-ME on the (a) benchmark and (b) BETA datasets with different numbers of training blocks. Error bars denote SEM and asterisks denote significant differences between "w/ EEG-ME" and "w/o EEG-ME" by paired t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

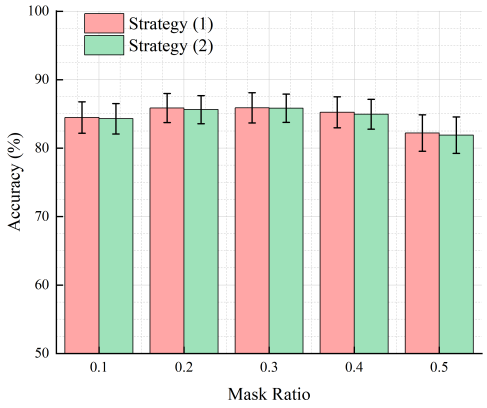

Fig. 8. Average classification accuracy of CNN-Former with two EEG-ME variation strategies across all subjects on the benchmark dataset with different mask ratios. Error bars denote SEM. P values between the two variation strategies by paired t-test are higher than 0.05 at all mask ratios.

with Bonferroni correction reveals that EEG-ME can significantly improve the classification performance of CNN-Former at different training blocks ($p < 0.001$). Generally, ΔP increases as the number of training blocks decreases.

Furthermore, comparative experiments are conducted with two variation strategies of EEG-ME, using a data length of 1 s. The benchmark dataset is utilized here due to its relatively high SNR. Fig. 8 illustrates the average classification accuracy of CNN-Former with the two strategies across all subjects on the benchmark dataset, with different mask ratios. The paired t-test reveals no significant difference between the two variation strategies at all mask ratios (all $p > 0.05$).

IV. Discussion

This study proposes EEG-ME as a novel data augmentation approach for DL-based asynchronous SSVEP classification algorithms. In this section, we analyze the factors influencing the performance of EEG-ME and the underlying theories behind the effectiveness of EEG-ME. In addition, we analyze the asynchronous SSVEP classification algorithms via frame-by-frame detection.

TABLE VI
$ratio_b$ OF EEG-ME-BASED DL METHODS WITH DIFFERENT DATA LENGTHS OF TIME WINDOWS ON THE BENCHMARK AND BETA DATASETS

		0.8 s	1 s	1.2 s
Benchmark	**CNN-Former**	0.3	0.3	0.3
	tCNN	0.1	0.2	0.2
	EEGNet	0.2	0.2	0.2
BETA	**CNN-Former**	0.3	0.3	0.3
	tCNN	0.2	0.2	0.2
	EEGNet	0.1	0.2	0.2

TABLE VII
PARAMETERS OF DL-BASED METHODS WITH DIFFERENT DATA LENGTHS OF TIME WINDOWS

	0.8 s	1 s	1.2 s
CNN-Former	1,253,768	1,509,768	1,765,768
tCNN	883,880	1,098,920	1,313,960
EEGNet	97,576	115,816	137,896

A. Analysis of Factors Influencing the Effectiveness of EEG-ME

In this subsection, we assess the influence of mask ratios, the number of original samples, and data lengths of time windows on the effectiveness of EEG-ME, as well as evaluate the effectiveness of EEG-ME on different DL-based methods.

Firstly, the effectiveness of EEG-ME is evaluated with respect to the influence of the mask ratio. Table III and Table IV indicate that the average classification accuracy reaches its optimal performance at a specific mask ratio ($ratio_b$), which represents a balance between the reduction of overfitting and the loss of information. A higher mask ratio beyond a certain point may result in significant information loss and prevent the model from learning effective features. Table VI illustrates the $ratio_b$ values for different DL-based methods with varying data lengths of time windows on the benchmark and BETA datasets. The $ratio_b$ values differ across DL-based methods, possibly due to the difference in the learning effectiveness of network architectures. An enhanced network architecture allows more information loss of training samples, thereby resulting in a higher $ratio_b$. Table VII shows the total parameters for DL-based methods, with CNN-Former having more parameters compared with the other two methods. The larger number of parameters and novel network architecture integrating CNN with Transformer result in CNN-Former having a better learning capacity, which may explain why its $ratio_b$ is larger than that of the other two methods. The proposed EEG-ME approach can improve the classification performance of DL-based methods with appropriate mask ratios. Based on Table III, Table IV, and Table VI, a mask ratio of 0.2 is recommended for general network architectures.

Furthermore, the effectiveness of EEG-ME is evaluated in terms of the influence of the number of original samples. The benchmark dataset and the BETA dataset differ in both the lasted stimulation time and the number of blocks, resulting in a different number of original samples. As shown in Table III, Table IV, and Table VIII, P_{base} of the BETA dataset is smaller than that of the benchmark dataset, while ΔP of the BETA dataset is higher in general. The results suggest

TABLE VIII
ΔP (%) OF DL-BASED METHODS WITH DIFFERENT DATA LENGTHS OF TIME WINDOWS ON THE BENCHMARK AND BETA DATASETS

		0.8 s	1 s	1.2 s
Benchmark	**CNN-Former**	3.65	3.18	2.59
	tCNN	0.76	1.42	1.73
	EEGNet	2.91	3.06	2.88
BETA	**CNN-Former**	8.04	11.09	13.98
	tCNN	1.72	3.12	4.63
	EEGNet	1.69	2.81	2.78

that the effectiveness of EEG-ME is more significant when dealing with a smaller number of original samples. However, differences in experimental conditions between datasets [31], [32] may affect conclusions. To further validate this, CNN-Former is trained with and without EEG-ME on both datasets using varied numbers of blocks. In Fig. 7, ΔP increases with the training blocks decrease in general. The conclusion aligns with the previous observations, emphasizing that EEG-ME demonstrates significant benefits, particularly for a small number of original samples. This could be attributed to the higher risk of overfitting with a reduced sample size while EEG-ME effectively mitigates overfitting.

In addition, we evaluate the influence of the data length of time windows on the effectiveness of EEG-ME. In Fig. 6 and Table VIII, ΔP generally increases as the data length of time windows increases. This can be attributed to the amplified task-related information present in the EEG data, which mitigates the adverse impact of information loss. Consequently, EEG-ME exhibits increased performance with an increase in the data length of time windows. However, with data length from 0.8 s to 1.2 s, ΔP of CNN-Former on the benchmark dataset decreases with an increase in the data length of time windows. This may be because CNN-Former has better learning capacity at longer time windows, resulting in higher accuracy. Therefore, further improvements in classification performance become more difficult at longer time windows. Additionally, we observe that $ratio_b$ decreases as the data length decreases, particularly for CNN-Former in the 0.2 s to 0.6 s range. This decrease can be attributed to a decline in task-related information as the data length decreases, leading to a smaller allowed mask ratio.

Finally, we evaluate the effectiveness of EEG-ME on different DL-based methods, each with a unique network architecture. CNN-Former integrates CNN with Transformer, tCNN is a standard CNN network architecture without pooling layers, and EEGNet is a lightweight network architecture. In Table IV, P_{base} of CNN-Former is smaller than that of tCNN and EEGNet. When dealing with a smaller number of original samples, CNN-Former without EEG-ME appears to be more severely overfitted than tCNN and EEGNet, possibly due to its larger number of parameters. However, EEG-ME can address the overfitting problem by generating various training samples to increase model robustness. ΔP of CNN-Former is higher than that of tCNN and EEGNet, resulting in a higher P_{best} for CNN-Former over tCNN and EEGNet. This highlights the innovative network architecture of CNN-Former, which combines CNN and Transformer with a larger number of parameters, exhibiting a higher upper bound of classification performance. Overall, EEG-ME is more effective for DL-based methods with superior learning capacity, especially when dealing with a limited number of original samples.

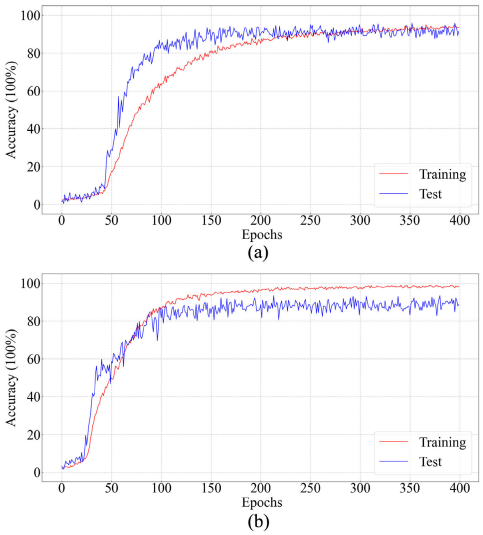

Fig. 9. Classification accuracy in the training and test sets during the training process of CNN-Former on a representative subject (a) with and (b) without EEG-ME. One epoch includes ten iterations.

B. Analysis of Underlying Theories Behind the Effectiveness of EEG-ME

This subsection examines the underlying theories of EEG-ME's contribution to the improved classification accuracy of SSVEP. Firstly, the effectiveness of EEG-ME in mitigating overfitting is demonstrated. Subsequently, potential theories explaining the effectiveness of EEG-ME are analyzed. The benchmark dataset is utilized due to its relatively high SNR, and the data length of time windows is set to 1 s. Two conditions, "w/ EEG-ME" and "w/o EEG-ME", are employed.

Overfitting is characterized by a model's strong performance on the training set but poor performance on the test set. The objective of mitigating overfitting is to prevent the model from learning the training data so well that it "remembers" the training data. To investigate whether EEG-ME mitigates overfitting, the classification accuracy between the training set and the test set is compared during the training process. The experiment is conducted on the first subject in the benchmark dataset. Fig. 9 illustrates the classification accuracy during the training process of CNN-Former with and without EEG-ME for the representative subject. The results demonstrate that EEG-ME effectively mitigates overfitting and improves the model's classification performance on the test set.

Furthermore, a comparative experiment is conducted to analyze the potential theories of EEG-ME effectiveness, using different training iterations. Fig. 10 illustrates the average classification accuracy of the CNN-Former with and without EEG-ME across all subjects on the benchmark dataset for varying training iterations. Before 1500 iterations, employing the paired t-test with Bonferroni correction reveals a significant superiority of the model without EEG-ME over the model

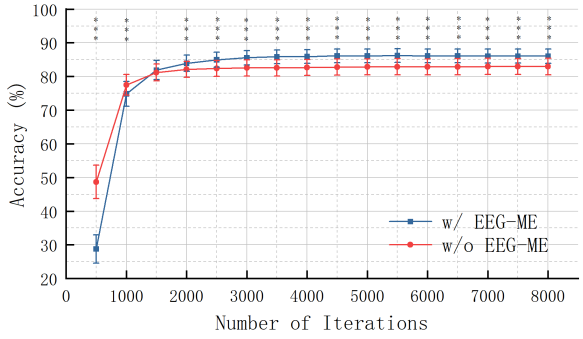

Fig. 10. Average classification accuracy of CNN-Former with and without EEG-ME using different numbers of training iterations. The error bars denote SEM and asterisks respectively denote significant differences between "w/ EEG-ME" and "w/o EEG-ME" by paired t-test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

with EEG-ME ($p < 0.001$). However, as the number of iterations increases, the paired t-test with Bonferroni correction consistently reveals a significant advantage of the model with EEG-ME over its counterpart without EEG-ME ($p < 0.001$). This may be attributed to EEG-ME challenging the feature learning process of the model, leading it to learn more robust features from the data.

In conclusion, EEG-ME masks partial data of the original EEG sample, hindering the model's ability to extract representations. This forces the model to learn more robust features from EEG data, potentially emphasizing the frequency feature of the SSVEP-EEG. This enables the model to better comprehend the underlying patterns in the EEG data rather than relying on memorization. Consequently, EEG-ME effectively mitigates overfitting and enhances the model's performance during testing. Essentially, EEG-ME enhances the model's generalization capabilities.

C. Feature Visualization

To provide further insight into how the proposed data augmentation approach improves network classification performance, feature visualization is conducted on the 1 s-based CNN-Former models. For enhanced clarity, the feature visualization is performed on the BETA dataset, as the proposed EEG-ME exhibited superior performance on this dataset. The output features of the first convolution layer, the multi-scale block, the last convolution layer, the first layer of the Transformer module, and the second layer of the Transformer module are visualized. The visualization process involved averaging the output features across dimensions and applying the Fast Fourier Transform (FFT) to obtain the amplitude spectrum. Subsequently, the amplitude spectrums are averaged across all subjects in the BETA dataset. It is important to note that, due to downsampling (step size of 5), the output units of each dimension of the last convolution layer only consisted of 50 points. It means that the FFT can only depict information within the 25 Hz range. The representative stimulus target with an integer Hz frequency is chosen for analysis. Fig. 11 presents the averaged amplitude spectrum of the output features from the representative layers of the CNN-Former model with and without the EEG-ME method across all subjects in the BETA dataset for each representative stimulus target. In general, the peak of the target frequency point for the "w/ EEG-ME" is more pronounced compared to the "w/o EEG-ME". This demonstrates that EEG-ME encourages the model to assign greater attention to the frequency characteristics of SSVEP-EEG.

D. Analysis of Asynchronous SSVEP Classification Algorithms

The asynchronous system requires continuous decoding of EEG signals using a fixed interval [9]. For example, a 1-second time window slides over the EEG data every 0.1 s. However, synchronizing the start point of the sliding time window with the generation of the SSVEP component is challenging in the asynchronous system. As a result, continuous decoding of EEG signals often occurs near the appearance of the SSVEP component, and each frame around this time may serve as a potential start point for the decoding time window.

This subsection explores the performance differences between asynchronous and synchronous SSVEP-EEG decoding algorithms in frame-by-frame detection. We use EEG data from the sixth block of the first subject in the benchmark dataset for detection analysis, while other blocks of this subject are reserved for training. The data length of time windows is fixed at 1 s and the start point of time windows slips 375 frames, including 0.5 s cue time and 1 s stimulation time. The detection results are obtained by averaging the test accuracy of 40 trials for each frame. We assume the SSVEP component appears at 0.14 s [31] after the stimulation onset.

DL-based methods for SSVEP classification are influenced by the selection way of training samples during the training phase. Synchronous algorithms typically use pre-defined time windows [19], [20], [21], such as $[0.14, 0.14 + d_s]$ s during stimulation time, for training, while asynchronous algorithms can employ a random sample selection strategy [7] as described in Section II-C. To emphasize the impact of training strategies on model performance, we use CNN-Former as the base model in both DL-based synchronous and asynchronous algorithms. Besides, TDCCA [14] and CCA [23] are respectively utilized as synchronous and asynchronous algorithms to evaluate the results of traditional methods.

Fig. 12(a) shows the frame-by-frame detection results of the DL-based methods, while Fig. 12(b) illustrates the results of the CCA-based methods. Both DL-based and CCA-based methods in the asynchronous algorithms exhibit consistent stability across all frames during the stimulation time. In contrast, the synchronous algorithms perform well at the appearance of the SSVEP component (0.14 s after the stimulation onset) but poorly in other frames. Furthermore, in real-life asynchronous SSVEP-BCI applications, the stimulus target is always in flickering mode. The stimulus target's phase information (initial brightness), which is crucial for synchronization algorithms [10], [11], cannot be effectively utilized. Hence, the practical performance of synchronization algorithms may be inferior to that observed in our simulations.

In summary, synchronous algorithms outperform asynchronous algorithms in predefined time windows due to their utilization of the phase information of the stimulus target

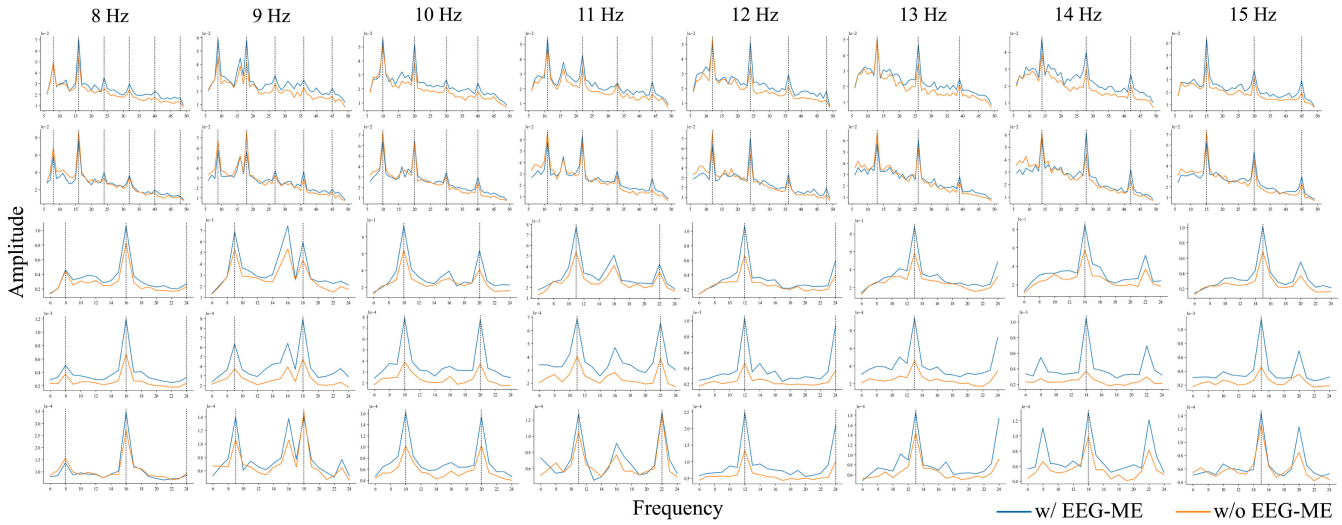

Fig. 11. Averaged amplitude spectrum of the output features of the first convolution layer (first row), the multi-scale block (second row), the last convolution layer (third row), the first layer of the Transformer module (fourth row), and the second layer of the Transformer module (fifth row) across all subjects of the BETA dataset for each representative stimulus target. The dashed line represents the corresponding SSVEP frequency of the stimulus target.

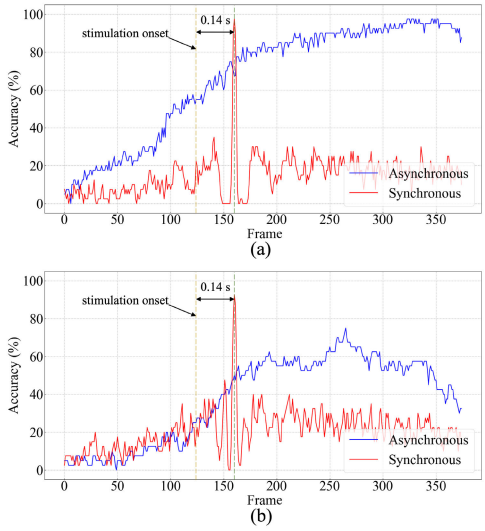

Fig. 12. Frame-by-frame analysis of (a) DL-based and (b) CCA-based methods.

[10], [11]. However, asynchronous algorithms demonstrate greater stability in decoding EEG data. Therefore, while synchronous algorithms may be suitable for simple tasks without system delay to build a high-speed synchronous BCI system, more robust asynchronous algorithms are necessary for asynchronous BCI systems.

V. Conclusion

This study proposes EEG-ME, a novel data augmentation method designed to address the challenge of the limited amount of EEG data in asynchronous SSVEP-BCI. EEG-ME is a parameter-free and easy-to-implement approach that can help models learn more robust features by masking partial EEG data, leading to enhanced generalization capabilities of models. As a result, it effectively enhances the classification performance of DL-based methods. Experiments conducted on benchmark and BETA datasets with various DL-based methods validate the effectiveness of EEG-ME. In future work, the proposed EEG-ME is promising to be combined with other data augmentation techniques and may inspire the development of novel data augmentation approaches in EEG and other physiological signal research fields.

References

[1] X. Pei, J. Hill, and G. Schalk, "Silent communication: Toward using brain signals," *IEEE Pulse*, vol. 3, no. 1, pp. 43–46, Jan. 2012.

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

[3] J. Jiang, C. Wang, J. Wu, W. Qin, M. Xu, and E. Yin, "Temporal combination pattern optimization based on feature selection method for motor imagery BCIs," *Frontiers Human Neurosci.*, vol. 14, p. 231, Jun. 2020.

[4] L. Xu, M. Xu, T.-P. Jung, and D. Ming, "Review of brain encoding and decoding mechanisms for EEG-based brain–computer interface," *Cognit. Neurodynamics*, vol. 15, no. 4, pp. 569–584, Aug. 2021.

[5] Y. Pei et al., "A tensor-based frequency features combination method for brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 465–475, 2022.

[6] L. Chen et al., "Adaptive asynchronous control system of robotic arm based on augmented reality-assisted brain–computer interface," *J. Neural Eng.*, vol. 18, no. 6, Dec. 2021, Art. no. 066005.

[7] W. Ding, J. Shan, B. Fang, C. Wang, F. Sun, and X. Li, "Filter bank convolutional neural network for short time-window steady-state visual evoked potential classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2615–2624, 2021.

[8] Y. Ke, J. Du, S. Liu, and D. Ming, "Enhancing detection of control state for high-speed asynchronous SSVEP-BCIs using frequency-specific framework," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1405–1417, 2023.

[9] G. Townsend, B. Graimann, and G. Pfurtscheller, "Continuous EEG classification during motor imagery-simulation of an asynchronous BCI," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 12, no. 2, pp. 258–265, Jun. 2004.

[10] 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.

[11] 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.

[12] Y. Zhang, G. Zhou, J. Jin, M. Wang, X. Wang, and A. Cichocki, "L1-regularized multiway canonical correlation analysis for SSVEP-based BCI," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 6, pp. 887–896, Nov. 2013.

[13] X. Yin and M. Lin, "Multi-information improves the performance of CCA-based SSVEP classification," *Cogn. Neurodynamics*, pp. 1–8, Jan. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s11571-022-09923-x

[14] Q. Wei, S. Zhu, Y. Wang, X. Gao, H. Guo, and X. Wu, "A training data-driven canonical correlation analysis algorithm for designing spatial filters to enhance performance of SSVEP-based BCIs," *Int. J. Neural Syst.*, vol. 30, no. 5, May 2020, Art. no. 2050020.

[15] 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.

[16] Q. Sun, M. Chen, L. Zhang, C. Li, and W. Kang, "Similarity-constrained task-related component analysis for enhancing SSVEP detection," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 046080.

[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] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[19] 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.

[20] J. Chen, Y. Zhang, Y. Pan, P. Xu, and C. Guan, "A transformer-based deep neural network model for SSVEP classification," *Neural Netw.*, vol. 164, pp. 521–534, Jul. 2023.

[21] X. Wang, A. Liu, L. Wu, C. Li, Y. Liu, and X. Chen, "A generalized zero-shot learning scheme for SSVEP-based BCI system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 863–874, 2023.

[22] B. Fang et al., "Brain–computer interface integrated with augmented reality for human–robot interaction," *IEEE Trans. Cognit. Develop. Syst.*, vol. 15, no. 4, pp. 1702–1711, Dec. 2023.

[23] 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. 54, no. 6, pp. 1172–1176, Jun. 2007.

[24] 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.

[25] 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.

[26] P. Saidi, A. Vosoughi, and G. Atia, "Detection of brain stimuli using Ramanujan periodicity transforms," *J. Neural Eng.*, vol. 16, no. 3, Jun. 2019, Art. no. 036021.

[27] X. Zhang et al., "A convolutional neural network for the detection of asynchronous steady state motion visual evoked potential," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1303–1311, Jun. 2019.

[28] A. Ravi, J. Lu, S. Pearce, and N. Jiang, "Enhanced system robustness of asynchronous BCI in augmented reality using steady-state motion visual evoked potential," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 85–95, 2022.

[29] 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.

[30] R. Luo, M. Xu, X. Zhou, X. Xiao, T.-P. Jung, and D. Ming, "Data augmentation of SSVEPs using source aliasing matrix estimation for brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 6, pp. 1775–1785, Jun. 2023.

[31] 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, Oct. 2017.

[32] 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.

[33] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *J. Neurosci. Methods*, vol. 346, Dec. 2020, Art. no. 108885.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] N. K. Nik Aznan, A. Atapour-Abarghouei, S. Bonner, J. D. Connolly, N. Al Moubayed, and T. P. Breckon, "Simulating brain signals: Creating synthetic EEG data via neural-based generative models for improved SSVEP classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[36] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.

[37] P. R. A. S. Bassi, W. Rampazzo, and R. Attux, "Transfer learning and SpecAugment applied to SSVEP based BCI classification," *Biomed. Signal Process. Control*, vol. 67, May 2021, Art. no. 102542.

[38] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[40] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[41] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2456–2467, Aug. 2022.

[42] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2023.

[43] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[44] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[47] Y. Pei et al., "Data augmentation: Using channel-level recombination to improve classification performance for motor imagery EEG," *Frontiers Human Neurosci.*, vol. 15, Mar. 2021, Art. no. 645952.