

# Functional Connectivity Guided Deep Neural Network for Decoding High-Level Visual Imagery

Byoung-Hee Kwon<sup>a</sup>, Minji Lee<sup>b</sup>, Seong-Whan Lee<sup>c,\*</sup>

<sup>a</sup>*Department of Brain and Cognitive Engineering, Korea University, Seoul, 02841, Republic of Korea*

<sup>b</sup>*Department of Biomedical Software Engineering, The Catholic University of Korea, Bucheon, 14662, Republic of Korea*

<sup>c</sup>*Department of Artificial Intelligence, Korea University, Seoul, 02841, Republic of Korea*

---

## Abstract

This study introduces a pioneering approach in brain-computer interface (BCI) technology, featuring our novel concept of high-level visual imagery for non-invasive electroencephalography (EEG)-based communication. High-level visual imagery, as proposed in our work, involves the user engaging in the mental visualization of complex upper limb movements. This innovative approach significantly enhances the BCI system, facilitating the extension of its applications to more sophisticated tasks such as EEG-based robotic arm control. By leveraging this advanced form of visual imagery, our study opens new horizons for intricate and intuitive mind-controlled interfaces. We developed an advanced deep learning architecture that integrates functional connectivity metrics with a convolutional neural network-image transformer. This framework is adept at decoding subtle user intentions, addressing the spatial variability in high-level visual tasks, and effectively translating these into precise commands for robotic arm control. Our comprehensive offline and pseudo-online evaluations demonstrate the framework's efficacy in real-time applications, including the nuanced control of robotic arms. The robustness of our approach is further validated through leave-one-subject-out cross-validation, marking a significant step towards versatile, subject-independent BCI applications. This research highlights the transformative impact of

---

\*Corresponding author.

Email addresses: [bh\\_kwon@korea.ac.kr](mailto:bh_kwon@korea.ac.kr) (Byoung-Hee Kwon),  
[minjilee@catholic.ac.kr](mailto:minjilee@catholic.ac.kr) (Minji Lee), [sw.lee@korea.ac.kr](mailto:sw.lee@korea.ac.kr) (Seong-Whan Lee)

advanced visual imagery and deep learning in enhancing the usability and adaptability of BCI systems, particularly in robotic arm manipulation.

*Keywords:*

Brain-computer interface, Deep learning, Functional connectivity,  
Electroencephalography, Visual imagery

---

## 1. Introduction

Brain-computer interfaces (BCIs) are an advanced technology that enables communication between machines and users, translating human brain signals into machine commands to reflect the user's intentions. Non-invasive BCI is a practical technology as no surgical operation is required (Kwak and Lee, 2019; Jeong et al., 2022; Sun et al., 2024). Electroencephalography (EEG) signals have the advantage of having better time resolution than comparable methods such as functional magnetic resonance imaging and near-infrared spectroscopy. Thus, EEG allows rapid communication between users and external devices which facilitates the technological improvement of rehabilitation systems for patients with tetraplegia or supporting the daily life activities of healthy people (Suk and Lee, 2012; Zhou et al., 2023; Ai et al., 2023). Using visual imagery, which is a type of intuitive BCI paradigm, users could control their avatars and robotic devices in virtual space and in the real-world environment by reflecting user intentions. Furthermore, it allows the creation of a new type of world in which communication in a 3D virtual space, without any spatial or physical restrictions, becomes possible by using only brain signals.

Previous BCI studies have focused on a variety of BCI paradigms to identify users' intent to move. Among the endogenous BCI paradigms to control BCI-based devices, motor imagery (MI) (Sun et al., 2019; C. I. Penaloza and S. Nishio, 2018; Jeong et al., 2022; Liu et al., 2024) is an effective option for controlling practical BCI applications such as robotic arms or wheelchairs. When a user performs MI, event-related desynchronization/synchronization (ERD/ERS), referred to as sensorimotor rhythm, is generated in the motor cortex (Jeong et al., 2020; Wang et al., 2020; Luo, 2024). ERD/ERS can be induced from the mu band ([8–12] Hz) and beta band ([13–30] Hz), respectively. However, despite its prevalence, the MI paradigm in BCI studies poses challenges in intuitiveness for users, especially when they imagine muscle movements. Although a user imagines the same muscle movement when

performing the MI paradigm, it is difficult to detect the user’s intentions because the time and sequence of imagery in which the user imagines muscle movements are not constant. In addition, performing the MI task is accepted differently by users, and this phenomenon prevents users from having a uniform imagination (Pérez-Velasco et al., 2022; Tibrewal et al., 2022). This leads to motion-related signals in different brain regions for each person and degrades the practicality of BCI applications.

Visual imagery is an endogenous BCI paradigm that allows users to use more intuitive imagination than MI when performing imagery tasks (Kwon et al., 2020), regardless of the complexity of the movement, reducing the difficulty of imagination and minimizing user fatigue. Additionally, it can be used to lead users to perform the visual imagery task uniformly. This effective BCI paradigm presents the possibility for the development of technologies that control avatars or robots by reflecting users’ intentions in 3D virtual reality and in the real-world by using only brain signals. To decode user intention, we used neural response patterns that included delta, theta, and alpha frequency bands. Brain activities based on visual imagery induce delta and theta bands in the prefrontal lobe and the alpha band in the occipital lobe, which includes the visual cortex (Pearson, 2019; Koizumi et al., 2019). Previous studies proved that the brain signals generated by visual imagery can be analyzed by specific brain regions and frequencies (Sousa et al., 2017) to control BCI-related devices.

However, most previous visual imagery paradigms for EEG-based BCI have focused on tasks designed in two-dimensional space, often lacking contextual relevance or real-world applicability. Common examples include imagining directional arrows pointing left or right (Sousa et al., 2017), mentally visualizing geometric shapes such as circles or triangles (Llorella et al., 2021), or recalling specific objects or colors (Xie et al., 2020; Kilmarx et al., 2024). While these tasks are easy to implement and elicit measurable brain responses, they often involve low task complexity and offer limited relevance to real-world BCI applications. Even in studies that extended visual imagery into three-dimensional space, the tasks were frequently limited to imagining spatial directions or simple trajectories (Koizumi et al., 2019), rather than simulating functional, goal-directed actions. As a result, such paradigms may not effectively support intuitive and naturalistic control of external devices. These limitations highlight the need for a more realistic and complex paradigm that better engages the brain’s visual and motor-related regions through structured and contextually grounded imagery tasks. Therefore,

this study proposes a high-level visual imagery paradigm that not only provides intuitive visual stimuli for users but also delivers complex movement cues required in virtual reality and real-world environments. In addition, it allows the training of a user by providing repeated 3D space-based visual stimuli. High-level visual imagery allows users to perform more advanced tasks when performing BCI-based device control because the user is given a moving stimulus.

In this study, visual imagery (Sousa et al., 2017; Koizumi et al., 2019) was used to provide more intuitive stimuli to users. We provided high-level visual imagery by which the user imagined complex movements, unlike traditional visual imagery by which the user imagined simple movements. Traditional visual imagery provides simple stimulation for imagining, such as static images or direction indications (Kwon et al., 2020). We provided visual stimuli in a three-dimensional space on the monitor to enable users to perform visual imagery effectively. The high-level visual imagery paradigm that we designed presented the high-complexity movement stimulation required in virtual reality and real-life to help the user intuitively perform imagery tasks. In addition, it allowed the training of a user by providing repeated 3D space-based visual stimuli. High-level visual imagery allows users to perform more advanced tasks when performing BCI-based device control because the user is given a moving stimulus.

This advanced form of mental imagery, involving the imagination of three-dimensional constructs, engages various brain regions in a sophisticated and coordinated manner. Initially, the occipital lobe, located in the posterior part of the brain, is essential in providing fundamental visual processing capabilities. Although the task is imaginative rather than perceptive, this region still plays a role in offering a base from previously acquired visual memories and experiences, crucial for constructing the initial framework of the imagined 3D image.

The frontal lobe, particularly the dorsolateral prefrontal cortex, then takes a leading role in the cognitive aspects of imagining these 3D constructs. It goes beyond mere visual processing to engage in higher-order cognitive functions. This includes the manipulation of abstract spatial concepts and the organization of complex visual information derived from memory and experience. The frontal lobe's executive functions are vital in this context, enabling the individual to mentally rotate, scale, and manipulate the imagined three-dimensional objects. Additionally, its working memory plays a crucial role in maintaining a dynamic and coherent representation of these imag-

ined constructs over time, allowing for the creation of sustained, detailed, and vivid mental imagery. In summary, while the occipital lobe provides the necessary visual references from memory, the frontal lobe's contribution is pivotal in transforming these references into intricate three-dimensional mental constructs. This process involves complex cognitive abilities like spatial reasoning, memory retrieval, and the imaginative manipulation of visual elements, underscoring the brain's remarkable capacity to not only perceive the external world but also to internally reconstruct and explore complex spatial environments within the mind's eye.

We proposed a deep learning framework based on functional connectivity and a convolutional neural network (CNN)-image transformer network to facilitate robust decoding of high-level visual imagery. We decoded brain signals derived from users when imagining four types of motions (e.g., picking up a cell phone, pouring water, opening a door, and eating food) and evaluated the performance of our network, which considers both spatial and temporal information.

The contributions of this study are threefold. (*i*) We proposed a functional connectivity-based deep neural network to decode high-level visual imagery from subjects. As a certain spatial pattern is found in visual imagery, our proposed network emphasizes spatial information based on the phase-locking value (PLV) for robust decoding performance. (*ii*) We measured the model performance through not only offline analysis but also pseudo-online analysis to confirm the feasibility of a real-time scenario for a practical BCI. (*iii*) In addition, to contribute to practical BCI application, the possibility of robust classification in the subject-independent condition was investigated by applying a leave-one-subject-out (LOSO) evaluation.

## 2. Materials and Methods

### 2.1. Participants

Fifteen subjects (Sub01–Sub15; aged 20–30 years) who had no neurological disease participated in our experiments. The subjects were asked to avoid anything that could affect their physical and mental condition during the experiment, such as drinking alcohol or using psychotropic drugs, during the day before the experiment. In addition, the participants were instructed to sleep for more than 8 h on the day before the experiment. This study was reviewed and approved by the Institutional Review Board at Korea University

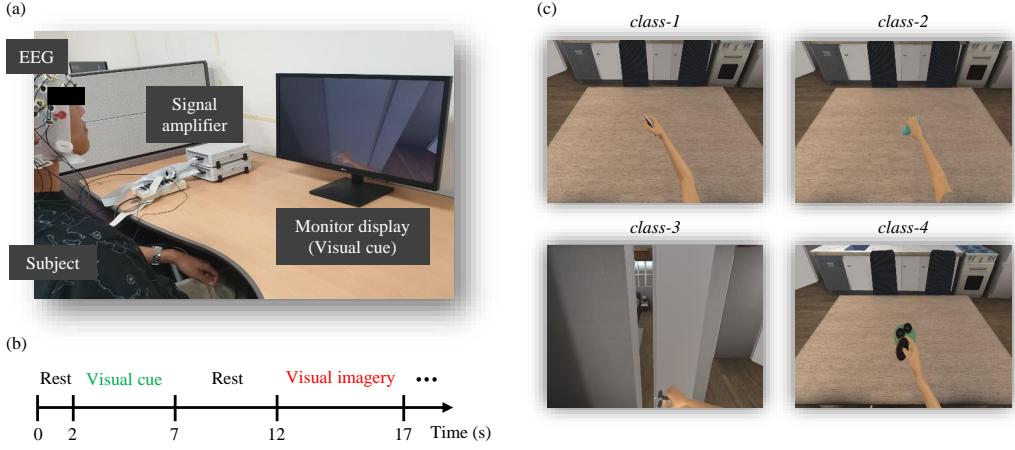


Figure 1: Experimental protocols for high-level visual imagery from electroencephalography (EEG) signals. (a) Experimental environment for acquiring high-level visual imagery data. (b) Experimental paradigm in a single trial and representation of visual cues according to each task. (c) The type of stimuli given to users in the experiment: picking up a cell phone (class-1), pouring water (class-2), opening a door (class-3), and eating food (class-4).

(KUIRB-2020-0013-01), and written informed consent was obtained from all participants before the experiment.

## 2.2. Data Acquisition

EEG data were recorded using an EEG signal amplifier (BrainAmp, Brain Products GmbH, Gilching, Germany) with MATLAB 2019a (MathWorks, Natick, MA, USA), sampled at 1,000 Hz. Additionally, we applied a 60 Hz notch filter to reduce the effect of external electrical noise (for example, direct-current noise owing to the power supply, scan rate of the monitor display, and frequency of the fluorescent lamp) in the raw signals. EEG was recorded from 64 Ag/AgCl electrodes according to the International 10–20 system (Fp1-2, AF5-6, AF7-8, AFz, F1-8, Fz, FT7-8, FC1-6, T7-8, C1-6, Cz, TP7-8, CP1-6, CPz, P1-8, Pz, PO3-4, PO7-8, POz, O1-2, Oz, and Iz). The ground and reference channels were placed on the Fpz and FCz, respectively. To maintain the impedance between the electrodes and skin below  $10\text{ k}\Omega$ , we injected conductive gel into the electrodes using a syringe with a blunt needle. A display monitor to provide the experimental paradigm to the subject was placed at a distance of approximately 90 cm to achieve a comfortable state

for the subject, and the subject sat in a comfortable position to conduct the experiment, as shown in Fig. 1(a).

### 2.3. Experimental Paradigm

As shown in Fig. 1(b), in this experimental paradigm, a single trial included four different phases. The first phase was a resting state during which a fixation cross was presented on the screen to provide a comfortable environment for the subjects for 2 s, before providing the visual stimuli. The visual stimuli that subjects should imagine were presented as a visual cue in the second phase. The third phase was another resting state, with the fixation cross presented, to provide sufficient time (5 s) to remove the afterimage of the stimulus presented in the second stage. This stage was essential for obtaining clear and meaningful EEG data from the visual imagery stage. In the fourth stage, after the 5-s resting phase, the subjects performed visual imagery for 5 s based on the visual cues provided during the second stage. During the visual imagery phase, the subjects saw a blank screen, with their eyes open, and imagined drawing a scene on the screen. In this process, the subject was asked to perform visual imagery with the feeling that he/she did not use any muscles, to avoid confusion with the MI. The duration of a single trial consisting of all four above-mentioned stages was 17 s. Each subject performed 50 trials in each class, for a total of 800 trials. Trials for each class were conducted in a random order across subjects.

We designed a training platform with moving visual stimulation, which we called 3D-BCI, as a guide to experimental protocols. These stimuli provide guidance on what the subject should imagine. The subjects performed high-level visual imagery based on a visual cue. We used 3D simulation software such as Unity 3D (Unity Technologies, San Francisco, CA, USA) and Blender (Blender 3D Engine: [www.blender.org](http://www.blender.org)) to design videos that were presented to the subjects. The visual stimuli consisted of four different scenarios: picking up a cell phone (*class-1*), pouring water (*class-2*), opening a door (*class-3*), and eating food (*class-4*). The classes were designed to reflect movements that the user could perform in real life, as well as those required for EEG-based robotic arm control.

### 2.4. Pre-processing

The EEG data were pre-processed using the BCI toolbox (Blankertz et al., 2010) in a MATLAB 2019a environment. Raw EEG data were down-sampled from 1,000 Hz to 250 Hz. We applied band-pass filters to extract

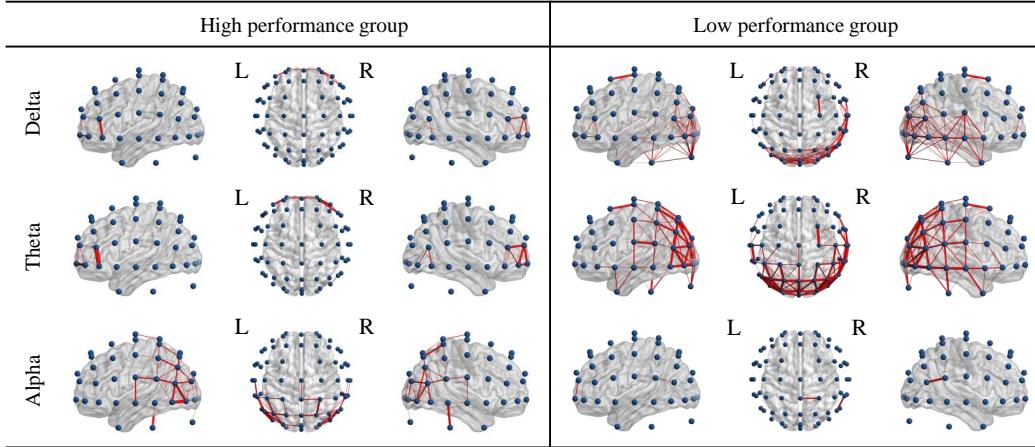


Figure 2: The connections between the channels that have functional connectivity scores above 0.9 when the high-performance group (Sub14, Sub09, and Sub04) and low-performance group (Sub01, Sub08, and Sub13) performed a high-level visual imagery task. Functional connectivity was assessed in the delta and alpha frequency ranges. In the delta band, connections are mainly represented in the prefrontal area as shown in the high-performance group. On the other hand, the low-performance group’s connections are irregular. In the alpha band, the third row, the high-performance group’s connections are mainly represented in the occipital area. However, the low-performance group’s connections show irregular tendencies.

specific frequency bands significant to visual imagery: delta [0.5-4] Hz, theta [4-8] Hz, and alpha [8-13] Hz, using Hamming-windowed zero-phase finite-impulse response filters with an optimized order ( $N = 30$ ) (Sousa et al., 2017; Koizumi et al., 2019). Opting for an advanced data augmentation technique, we integrated Gaussian noise into our dataset, a method renowned for its efficacy in enhancing data robustness and diversity. Originally, each subject’s dataset comprised 800 samples. With the application of our Gaussian noise-based augmentation, we realized a substantial fivefold increase, culminating in 4,000 samples per subject, thereby significantly expanding the training data’s dimensional space. With this enriched dataset, the data was systematically partitioned: 60% served as the foundational training dataset, 20% was delineated for validation, and the residual 20% was allocated for testing.

### 2.5. Functional Connectivity-based Deep Neural Network

Functional connectivity is used to assess neural interactions (Honey et al., 2009). We utilized functional connectivity as a method to evaluate regional

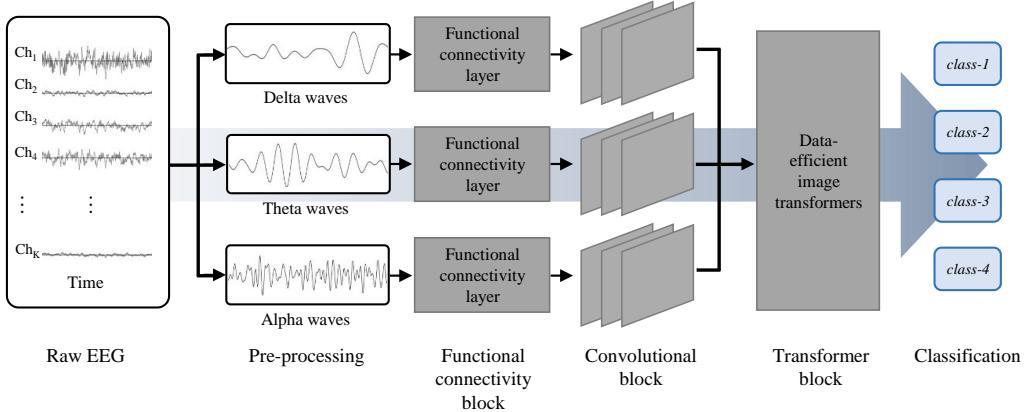


Figure 3: The overview of the proposed architecture begins with the segmentation of the raw electroencephalogram (EEG) into delta waves, theta waves, and alpha waves. In the functional connectivity layer, each of these EEG wave types is then multiplied channel-wise using a functional connectivity score measured by the phase-locking value. The modified EEG data that emphasize spatial information is used as the input data for the deep learning architecture. We extracted temporal information using a network consisting of a convolutional block, followed by spatial information via a transformer block.

interactions in the brain when a subject performs visual imagery. We selected the PLV method (Lachaux et al., 1999; Roach and Mathalon, 2008) as a functional connectivity method to calculate the relation score of each channel.

Brain signals can be expressed through correlations between brain regions. To decode user intention, raw EEG signals, which do not reflect correlations between regions of the brain and the deep neural network, lead to performance degradation. Therefore, we measured the functional connectivity scores using PLVs to consider spatial relationships. As shown in Fig. 2, we visualized the functional connectivity when the subjects performed the high-level visual imagery in each class, by using chord diagrams to confirm if this method was appropriate for decoding high-level visual imagery. In this study, subjects were categorized into high-performance and low-performance groups according to the classification performance criteria detailed in the results section. We identified the functional connectivity in all classes using the delta ([0.5–4] Hz), theta ([4–8] Hz), and alpha ([8–13] Hz) frequency ranges where significant features appear in high-level visual imagery. To extract signals from each frequency band, we employed the following generalized formula:

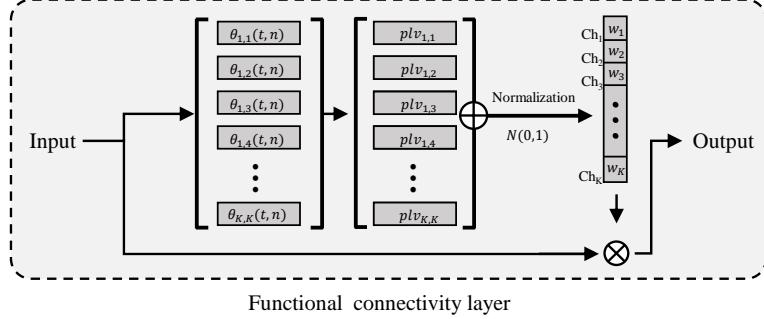


Figure 4: In the functional connectivity layer, the raw electroencephalogram (EEG) data undergo transformation through the phase locking value (PLV). Each channel’s connectivity score, derived via the PLV, is then normalized. The subsequent integration with the original EEG accentuates spatial details, producing an EEG output enriched with heightened spatial significance.

$$X_{\Delta f}(t) = \int_{f_{\min}}^{f_{\max}} X(f) e^{2\pi i f t} df \quad (1)$$

In this formula,  $X_{\Delta f}(t)$  represents the signal in the time domain for a specific frequency band,  $f_{\min}$  and  $f_{\max}$  define the lower and upper limits of the frequency band, and  $X(f)$  is the frequency spectrum of the original EEG signal. In the high-performance group (Sub14, Sub09, and Sub04), strong correlations were observed primarily near the prefrontal lobe in the delta and theta regions, and in the alpha frequency range, predominantly near the occipital lobe.

On the other hand, the low-performance group (Sub01, Sub08, and Sub13) demonstrated weak connections in the prefrontal lobe and occipital lobe. These results demonstrated that applying PLVs to decode high-level visual imagery is reasonable.

In this study, we proposed a functional connectivity-guided deep neural network (FCDN) to decode acquired high-level visual imagery data. The overall processing pipeline of the proposed FCDN begins with frequency-wise decomposition of the raw EEG signals, followed by spatial enhancement using functional connectivity scores. The resulting signals are then passed through a convolutional block to extract temporal features, which are subsequently transformed into spatial feature maps and processed by a data-efficient transformer for final classification. An overview of the proposed

Table 1: Description of the Proposed Architecture

<b>Layer</b>	<b>Type</b>	<b>Parameter</b>	<b>Output size</b>
1	Input	-	$1 \times 64 \times 1000$
2	Convolution	Filter size: $1 \times 20$	
		Stride size: $1 \times 1$	
		Feature map: 40	$40 \times 64 \times 981$
3	BatchNorm	-	
	Convolution	Filter size: $1 \times 20$	
		Stride size: $1 \times 1$	
		Feature map: 40	$80 \times 64 \times 962$
4	BatchNorm	-	
	Activation(ELU)	-	
		-	
5	Convolution	Filter size: $1 \times 40$	
		Stride size: $1 \times 1$	
		Feature map: 80	$160 \times 64 \times 962$
		-	
6	BatchNorm	-	
		-	
		-	
7	Activation(ELU)	Filter size: $1 \times 32$	
		Stride size: $1 \times 32$	
		Dropout(0.5)	$160 \times 64 \times 30$
		-	
8	Average pooling	Filter size: $1 \times 30$	
		Stride size: $1 \times 30$	$160 \times 64 \times 1$
		Dropout(0.5)	
9	Interpolation	-	$224 \times 224$
8	Concatenation	-	$224 \times 224 \times 3$
9	DeiT	-	$1 \times 4$

FCDN is presented in Fig. 3, and the architecture design of the deep learning part is presented in Table I. As a first step of the proposed network, raw EEG signals were decomposed into distinct frequency bands, specifically delta, theta, and alpha. These segregated bands were then utilized as inputs to their respective functional connectivity layers, serving as an input to the subsequent convolutional processing. A detailed representation of the functional connectivity layer is illustrated in Fig. 4, where the specific format of this layer is mathematically denoted as

$$PLV = \{plv_{k_1, k_2} \mid 1 \leq k_1 \leq K, 1 \leq k_2 \leq K\}, \quad (2)$$

where  $k_1$  and  $k_2$  are the channels in which we compare the phase, and  $K$  is the number of electrodes used. The phase signal  $\phi_{k_1}(t, n)$  is extracted for all time bins  $t$  and trials  $n$   $[1, \dots, N]$ , and the phase difference between  $k_1$  and  $k_2$  is derived from  $\theta(t, n) = \phi_{k_1}(t, n) - \phi_{k_2}(t, n)$ .

$$plv_{k_1, k_2} = \frac{1}{NT} \sum_{t=1}^T \sum_{n=1}^N e^{j\theta(t, n)} \quad (3)$$

These formulas can be used to understand the correlation between channels and to emphasize the spatial information in raw EEG signals.  $plv_{k_1, k_2}$  obtained from the above formula is an upper triangular matrix, and it was transposed to create a lower triangular matrix and added to the two matrices to create a new matrix  $S_{k_1 \text{and}, k_2}$  in the form of a symmetric matrix.

$$S_{k_1, k_2} = plv_{k_1, k_2} + (plv_{k_1, k_2})^T \quad (4)$$

To emphasize the spatial information in the input signal, we transformed the previously obtained  $S_{k_1, k_2}$  into a matrix of  $1 \times K$ , using the following formula:

$$\widetilde{PLV} = \sum_{k_1=1}^K S_{k_1, k_2} \quad (5)$$

We limited the distribution of  $\widetilde{PLV}$  between 0 and 1 using the simplest method, min-max normalization, to emphasize channels using  $\widetilde{PLV}$  implemented in one dimension.

$$w_K = \frac{\widetilde{PLV}_K - \widetilde{PLV}_{min}}{\widetilde{PLV}_{max} - \widetilde{PLV}_{min}}. \quad (6)$$

Consequently, for the 64 channels, the values of the channels that were highly correlated with the visual imagery approached 1, while that of channels that were not highly correlated with the high-level visual imagery had values that approached zero.  $w_K$  is a weight that emphasizes spatial information using channels related to visual imagery in the proposed network. The channel axis of the pre-processed EEG signal is multiplied by  $w_K$ , and the data are reconstructed according to the degree of visual imagery.

Conventional EEG-related deep learning frameworks use CNNs to train temporal features (Zhang et al., 2018; Amin et al., 2019; Zhang et al., 2019) and use image transformers to train spatial features (Mulkey et al., 2023). Based on these studies, we adopted the CNN framework in our proposed network to extract the information of spatially focused data using PLVs. We trained the network for 200 epochs, the batch size of the training process was set to 16, and the learning rate was set at 0.0001. When features are extracted from reconstructed EEG data using PLVs, the effect of emphasizing the channel that is most affected when performing visual imagery is greater than that of extracting features from raw EEG data.

The convolution block was designed to handle all electrode channels and to consider the temporal information of the reconstructed data. During this stage, each layer extracted the information that contained temporal information. Batch normalization was applied to standardize the variance of the learned data. Because the first convolutional layer did not improve performance even with nonlinear activation, we designed the network to maintain linearity without using the activation function in the first layer. In the second and third convolutional layers, we applied exponential linear units (ELUs) (Clevert et al., 2015) as an activation function. After the three convolutional layers, two average pooling layers were applied to down-sample the feature maps, primarily focusing on the extraction of temporal information. This restructuring to a 2D format was critical for subsequent integration with the data-efficient image transformers (DeiT) (Touvron et al., 2021), facilitating optimal capture of spatial information. In addition, we set the dropout probability to 0.5, after each average pooling layer, to help prevent the overfitting problem that occurred during training on small sample sizes. Then, we used bicubic interpolation (Dong et al., 2015) to resize the output of the convolution block.

$$F_{bicubic}(i, j) = \sum_{m=-1}^2 \sum_{n=-1}^2 a_{mn} \cdot (i - x_0)^m \cdot (j - y_0)^n \quad (7)$$

where  $F_{bicubic}$  is the feature map obtained after applying bicubic interpolation, where  $i$  and  $j$  denote the indices of rows and columns in the new feature map, respectively. The variables  $x_0$  and  $y_0$  correspond to the row and column indices of the original feature map, and  $a$  signifies the coefficients used in bicubic interpolation, computed using the neighboring pixel values of the

corresponding pixel in the original image. These coefficients must be pre-computed, as they are employed to fit a cubic polynomial. Additionally,  $m$  and  $n$  are indices ranging from -1 to 2, representing a  $4 \times 4$  pixel block. This facilitates utilization of the 16 pixels surrounding each pixel, to compute the new pixel value in the interpolated image.

After the bicubic interpolation, features derived from the distinct frequency bands—delta, theta, and alpha—are resized to a dimension of  $224 \times 224$ . Prior to further processing, these values are systematically normalized to fit within a 0 to 255 scale. This normalization step ensures the resultant feature map aligns with typical RGB channel conventions, thereby preparing the data for optimal compatibility with the subsequent deep learning architecture. Following this normalization, features from the three frequency domains are concatenated, resulting in a combined feature map of  $224 \times 224 \times 3$ . This procedure yields a comprehensive representation, integrating spectral EEG characteristics across the examined frequency bands. This unified  $224 \times 224 \times 3$  feature map is then input into the DeiT. Within the DeiT architecture, the combined EEG features are subjected to a series of convolutional and transformer-based operations, allowing for intricate spatial pattern recognition and further refinement.

In the realm of EEG signal processing, the scarcity of available data often poses significant challenges to conventional deep learning models. This limitation hampers the ability to effectively train large-scale models, thereby restricting the extraction of intricate and subtle features inherent in EEG signals. To mitigate this problem, our approach leverages the DeiT, a model specifically tailored to perform well with limited data.

The cornerstone of DeiT that makes it apt for handling limited data is the employment of a teacher-student distillation process. The key equation governing this distillation is as follows:

$$L_{distill} = \alpha \cdot L_{cls} + \beta \cdot \sum_{i=1}^N [sim(T_i, S_i) \cdot q_i] \quad (8)$$

where  $L_{distill}$  is the total distillation loss,  $L_{cls}$  is the classification loss,  $\alpha$  and  $\beta$  are hyperparameters controlling the balance between the two terms,  $T_i$  and  $S_i$  are the teacher's and student's hidden representations at layer  $i$ ,  $sim(\cdot, \cdot)$  is the similarity function, and  $q_i$  is the corresponding student's attention map.

By leveraging the distinctive information present in the EEG data’s delta, theta, and alpha waves, we construct three individual feature maps of dimension  $224 \times 224$ . Each of these maps has been created by undergoing processing through a functional connectivity block and a convolution block, with intricate features representative of their respective waveforms being extracted in the process. By stacking these processed maps along a third dimension, we obtain a composite representation of shape  $224 \times 224 \times 3$ . This method allows for an integrated perspective of the EEG signals, encapsulating the significant information and interactions of the delta, theta, and alpha frequencies.

Such a concatenated representation is seamlessly integrated with the DeiT model. This model architecture, which orchestrates a similarity loss between the hidden representations of a substantial pre-trained teacher model and a more compact student counterpart, guides the student to emulate the teacher’s behavior. This alignment capitalizes on the comprehensive knowledge contained within the teacher model, eliminating the necessity for extensive data. Furthermore, this mechanism enables DeiT to discern patterns across different regions of the  $224 \times 224 \times 3$  feature maps, adeptly harnessing spatial relationships and thereby achieving nuanced image understanding. The inclusion of a self-attention formulation further ensures the capture of complex dependencies, which empowers our model for this specific task. As a concluding step, the extracted features are introduced to a softmax classifier, with the categorical cross-entropy loss function and the Adam optimizer employed for training purposes.

### 3. Results

#### 3.1. Data Analysis

We used the BBCI toolbox and EEGLab toolbox (Delorme and Makeig, 2004) (version 14.1.2b) to verify the quality of the collected EEG data. The BBCI toolbox was used to measure the magnitude of the power of the high-level visual imagery data for each frequency range in each area of the brain. We used high-level visual imagery data for each stimulus and adopted a period of 0~5 s for the entire interval of the visual imagery phase. Among the 64 channels, Fz, representing the prefrontal lobe, and Oz, representing the occipital lobe, were selected to measure the power spectral changes (Wang et al., 2012; Carrier et al., 2001) from 0.1 Hz to 60 Hz in each selected channel. The frequency-specific power measured at each channel determines the area of the brain, which is significant when performing high-level visual imagery.

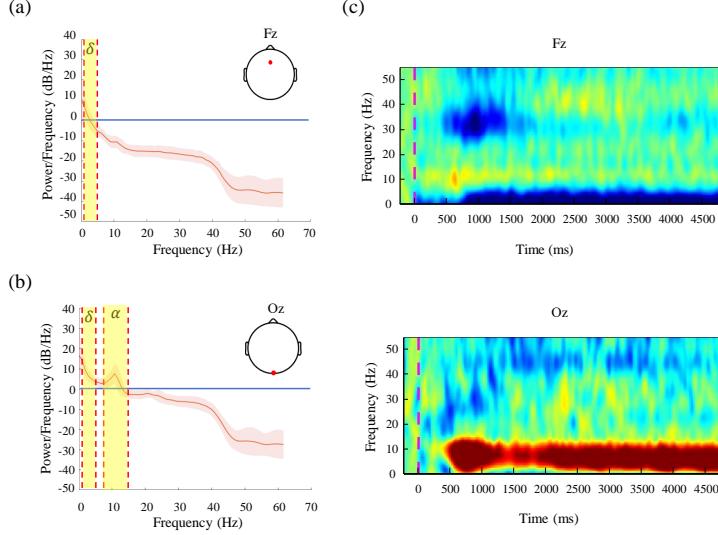


Figure 5: The power spectral changes in each selected channel. The yellow box in the graph indicates the frequency range containing a significant peak. The peaks were observed in channels (a) Fz and (b) Oz related to visual imagery. The time-frequency analysis (c) presents the distinct characteristics of high-level visual imagery. Prominent features are evident in the alpha band at the Fz and Oz locations, which are known to be associated with visual imagery.

While the user performed imagery, brain signal power changed significantly at channels representing the prefrontal and occipital lobes. A delta power peak was observed at the Fz channel (Fig. 5(a)), while delta power and alpha power peaks were found at the Oz channel (Fig. 5(b)).

Subsequently, we measured the variation in the spectral power of visual imagery based on the event-related spectral perturbation (ERSP) method (Makeig, 1993; Delorme and Makeig, 2004) on two channels that are considered to be related to visual imagery. ERSP analysis was performed between 0.5-50 Hz, using 400 time points. The baseline was set from -500 to 0 ms before the visual imagery phase in order to analyze the phenomena when subjects imagined an action. From the above results, it can be inferred that high-level visual imagery exhibits strong brain-signal features in the prefrontal and occipital lobes. Accordingly, we measured ERSP on Fz and Oz, two channels related to the visual imagery mentioned above. As shown in Fig. 5(c), Fz, which represents the prefrontal lobe, shows a strong activity power spectrum in the delta wave, and Oz, which represents the occipital

lobe, shows a strong activity power spectrum in the alpha wave.

### 3.2. Comparison of Classification Performance using FCDN and Baseline Methods

Table II compares the classification performance of the FCDN and conventional decoding methods. We performed five-fold cross-validation under a subject-dependent scheme to evaluate the classification performance of the proposed FCDN and baseline methods. We used the filter bank common spatial pattern (FBCSP) (Ang et al., 2008), EEGNet (Lawhern et al., 2018), ConvNet (Schirrmeister et al., 2017), FBCNet (Mane et al., 2021), and TSformer (Ahn et al., 2022) as baseline methods to decode high-level visual imagery data. The FBCSP methodology proficiently extracts spatial patterns across a spectrum of frequency bands, contributing significantly to the analysis of EEG signals. While EEGNet and ConvNet, as CNN-based models, emphasize local features due to their convolutional layers. While this approach of using convolutional layers in EEGNet and ConvNet has its merits, it might restrict their capability to encompass the wider spectrum of EEG signal characteristics. To address this, FBCNet integrates CNNs with handcrafted features, aiming to balance the representation of EEG signal attributes. Additionally, the TSformer, employing transformer techniques, primarily focuses on temporal dynamics, thereby achieving high performance. This method illustrates the potential of advanced signal processing techniques in enhancing EEG signal analysis. We computed a non-parametric paired permutation test (Maris and Oostenveld, Aug. 2007) to confirm the statistically significant differences between the classification results of the proposed method and the baseline methods. In this study, the average classification performance using the proposed network was 0.7234. Comparative analysis with other models showed that EEGNet achieved a performance of 0.6209, ConvNet reached 0.6447, FBCNet scored 0.6552, and TSformer attained 0.7072. The detailed classification performances under 2-class and 3-class conditions are presented in Supplementary Table II.

### 3.3. Ablation Study

In this study, we proposed FCDN, a functional connectivity-guided deep neural network designed for robust classification of EEG signal variability. To evaluate the contribution of the functional connectivity block, we visualized the output features at each major stage of the network using the t-distributed

Table 2: Comparison of Visual Imagery Classification Performances with Conventional Methods

<b>Subject</b>	<b>Method</b>					
	FBCSP	EEGNet	ConvNet	FBCNet	TSformer	FCDN
Sub01	0.4395 ( $\pm 0.0127$ )	0.5400 ( $+0.0252$ )	0.5650 ( $+0.0129$ )	0.5664 ( $\pm 0.0154$ )	0.6200 ( $\pm 0.0260$ )	<b>0.6273</b> ( $\pm 0.0123$ )
Sub02	0.4115 ( $\pm 0.0209$ )	0.5625 ( $+0.0362$ )	0.6113 ( $+0.0529$ )	0.6143 ( $\pm 0.0246$ )	0.6639 ( $\pm 0.0218$ )	<b>0.7135</b> ( $\pm 0.0195$ )
Sub03	0.4013 ( $\pm 0.0173$ )	0.5988 ( $+0.0195$ )	0.6248 ( $+0.0276$ )	0.6386 ( $\pm 0.0152$ )	<b>0.6894</b> ( $\pm 0.0179$ )	0.6738 ( $\pm 0.0207$ )
Sub04	0.5303 ( $\pm 0.0105$ )	0.6900 ( $\pm 0.0412$ )	0.6888 ( $\pm 0.0266$ )	0.7121 ( $\pm 0.0148$ )	0.7665 ( $\pm 0.0136$ )	<b>0.7701</b> ( $\pm 0.0214$ )
Sub05	0.5982 ( $\pm 0.0216$ )	0.5800 ( $\pm 0.0139$ )	0.6038 ( $\pm 0.0404$ )	0.6428 ( $\pm 0.0250$ )	<b>0.6938</b> ( $\pm 0.0163$ )	0.6871 ( $\pm 0.0186$ )
Sub06	0.5939 ( $\pm 0.0298$ )	0.5850 ( $\pm 0.0396$ )	0.6100 ( $\pm 0.0252$ )	0.6537 ( $\pm 0.0229$ )	0.7052 ( $\pm 0.0212$ )	<b>0.7203</b> ( $\pm 0.0188$ )
Sub07	0.5205 ( $\pm 0.0237$ )	0.5888 ( $\pm 0.0266$ )	0.6175 ( $\pm 0.0195$ )	0.6142 ( $\pm 0.0231$ )	0.6638 ( $\pm 0.0240$ )	<b>0.6813</b> ( $\pm 0.0239$ )
Sub08	0.4005 ( $\pm 0.0239$ )	0.5450 ( $\pm 0.0183$ )	0.5975 ( $\pm 0.0459$ )	0.6297 ( $\pm 0.0183$ )	0.6801 ( $\pm 0.0105$ )	<b>0.7195</b> ( $\pm 0.0267$ )
Sub09	0.3864 ( $\pm 0.0291$ )	0.7750 ( $\pm 0.0230$ )	0.7650 ( $\pm 0.0364$ )	0.6483 ( $\pm 0.0257$ )	0.6996 ( $\pm 0.0183$ )	<b>0.8207</b> ( $\pm 0.0196$ )
Sub10	0.4782 ( $\pm 0.0121$ )	0.5813 ( $\pm 0.0468$ )	0.6288 ( $\pm 0.0200$ )	0.6809 ( $\pm 0.0273$ )	<b>0.7338</b> ( $\pm 0.0198$ )	0.7118 ( $\pm 0.0265$ )
Sub11	0.5293 ( $\pm 0.0293$ )	0.5263 ( $\pm 0.0108$ )	0.5775 ( $\pm 0.0303$ )	0.6685 ( $\pm 0.0229$ )	<b>0.7208</b> ( $\pm 0.0242$ )	0.6601 ( $\pm 0.0105$ )
Sub12	0.4229 ( $\pm 0.0116$ )	0.6488 ( $\pm 0.0174$ )	0.6275 ( $\pm 0.0261$ )	0.6316 ( $\pm 0.0188$ )	0.6820 ( $\pm 0.0239$ )	<b>0.6951</b> ( $\pm 0.0177$ )
Sub13	0.5231 ( $\pm 0.0231$ )	0.5850 ( $\pm 0.0246$ )	0.5813 ( $\pm 0.0220$ )	0.6039 ( $\pm 0.0225$ )	0.6531 ( $\pm 0.0113$ )	<b>0.7203</b> ( $\pm 0.0190$ )
Sub14	0.4360 ( $\pm 0.0235$ )	0.7950 ( $\pm 0.0325$ )	0.8438 ( $\pm 0.0205$ )	0.8046 ( $\pm 0.0214$ )	0.8636 ( $\pm 0.0149$ )	<b>0.8720</b> ( $\pm 0.0139$ )
Sub15	0.5892 ( $\pm 0.0129$ )	0.7125 ( $\pm 0.0240$ )	0.7288 ( $\pm 0.0233$ )	0.7179 ( $\pm 0.0104$ )	0.7726 ( $\pm 0.0188$ )	<b>0.7785</b> ( $\pm 0.0264$ )
Avg.	0.4765 ( $\pm 0.0926$ )	0.6209 ( $\pm 0.0845$ )	0.6447 ( $\pm 0.0783$ )	0.6552 ( $\pm 0.1013$ )	0.7072 ( $\pm 0.0645$ )	<b>0.7234</b> ( $\pm 0.0614$ )
<i>p</i>	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

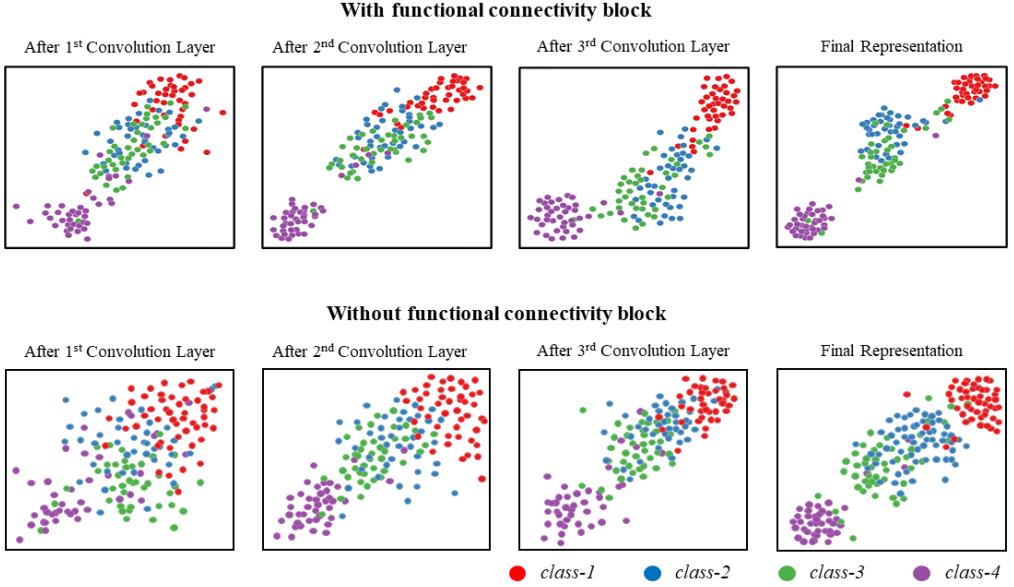


Figure 6: Representation of visualized feature distributions in each network layer using t-distributed stochastic neighbor embedding, with and without the functional connectivity block. Each column corresponds to the output after the 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> convolutional layers, and the final representation after the DeiT module. The top row shows the full FCDN model with the functional connectivity block, while the bottom row presents the ablation model without the functional connectivity block. As the network deepens, class-wise feature clustering becomes progressively clearer. Notably, the model with the functional connectivity block produces more distinct and compact class clusters compared to the model without the functional connectivity block, demonstrating the critical role of functional connectivity in enhancing spatial discriminability for high-level visual imagery decoding.

stochastic neighbor embedding (t-SNE) method (Jebelli et al., 2018; Arsalan et al., 2019).

Fig. 6 presents the t-SNE distributions of test-set features extracted after the 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> convolutional layers as well as the final representation from the DeiT module. The top row illustrates the feature representations obtained from the full FCDN model that includes the functional connectivity block, whereas the bottom row shows the results from the ablation model in which the functional connectivity block is removed. This structure allows for a direct visual comparison of the impact of the functional connectivity block across network depth. As the network becomes deeper, the feature clustering

Table 3: Comparison of Classification Accuracies With and Without the Functional Connectivity Block

<b>Subject</b>	<b>With</b>	<b>Without</b>	<b><math>\Delta</math> Accuracy</b>
Sub01	0.6273	0.6005	0.0268
Sub02	0.7135	0.6571	0.0564
Sub03	0.6738	0.6427	0.0311
Sub04	0.7701	0.7652	0.0049
Sub05	0.6871	0.6575	0.0296
Sub06	0.7203	0.6996	0.0207
Sub07	0.6813	0.6554	0.0259
Sub08	0.7195	0.7048	0.0147
Sub09	0.8207	0.7911	0.0296
Sub10	0.7118	0.6859	0.0259
Sub11	0.6601	0.6381	0.0220
Sub12	0.6951	0.7037	-0.0086
Sub13	0.7203	0.6845	0.0358
Sub14	0.8720	0.8467	0.0253
Sub15	0.7758	0.7623	0.0135
Avg.	0.7234	0.6997	0.0237

becomes more distinct. Notably, the model with the functional connectivity block produces tighter and more separable class clusters compared to the model without it. These observations suggest that the functional connectivity block enhances the spatial discriminability of neural representations during high-level visual imagery decoding.

To quantitatively validate these findings, we compared the classification performance of the full FCDN model with that of the ablated model. Table III reports the classification accuracies for each subject in both settings. The model including the functional connectivity block achieved better performance in 14 out of 15 subjects, resulting in an average accuracy improvement

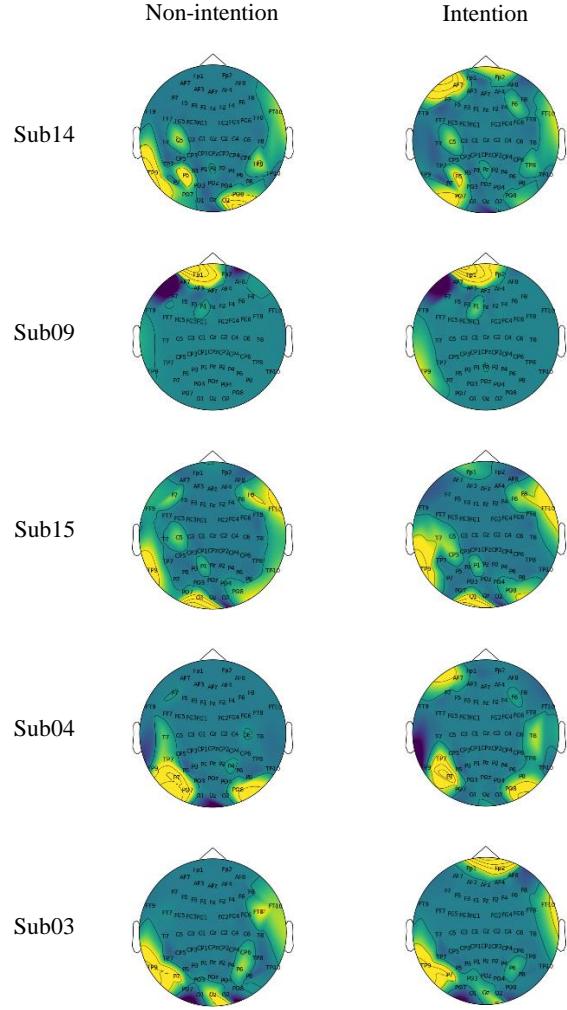


Figure 7: The figure depicts the outcomes of the top-five participants with high classification results, as processed through the convolution block, both without (Non-intention) and with (Intention) the application of functional connectivity. Remarkably, in all five participants, a strong correlation was observed in the frontal and occipital areas under the Intention condition.

of 2.37%. These results demonstrate that although the base model without the functional connectivity block is already well aligned with the characteristics of high-level visual imagery and captures relevant spatial-temporal features, the integration of visual cortex-guided functional connectivity fur-

ther contributes to enhancing decoding performance by improving the spatial discriminability of neural representations.

In Fig. 7, we elucidate the efficacy of the functional connectivity block by contrasting the feature maps post convolution block, both with and without the inclusion of the functional connectivity block. These feature maps are visualized using topographical plots. For a robust evaluation, the results are presented specifically for the top 5 subjects who demonstrated superior classification performance when processed through the FCDN. Notably, in the feature maps corresponding to the ‘Intention’ configuration (with the functional connectivity block), there is a discernible heightened activity in both the prefrontal and posterior regions, thereby underlining the pivotal role of the functional connectivity block. Conversely, the ‘Non-intention’ configuration (absent the functional connectivity block) exhibits a diminished neural representation in these critical regions.

To further validate the role of visual cortex-related neural information in decoding high-level visual imagery, we conducted a supplementary quantitative analysis by removing the occipital channels (e.g., PO3-4, PO7-8, POz, O1-2, Oz, and Iz) and re-evaluating classification performance. As shown in Supplementary Table I, removing the occipital region led to a considerable decline in accuracy across all subjects, with an average drop of 15.21%. This confirms that the posterior brain regions, particularly the occipital lobe associated with visual processing, play a crucial role in supporting the discriminative capacity of the proposed network.

### 3.4. Pseudo-Online Analysis

We evaluated the proposed method through a pseudo-online analysis (Jeong et al., 2020; Han et al., 2020) as depicted in Table IV. The EEG data were evaluated with a sliding window length of 2 s with a 50% overlap. For each window, the results of the predicted class were derived, and the final prediction results of the trial were determined using the averages of the results of four windows. We derived the results by constructing the datasets into three online runs for accurate verification. Fig. 8 shows the pseudo-online performance for the top-five subjects. *class-1* and *class-4* generally recorded high success rates across the entire class, which are consistent with the results shown in Fig. 6. The results for the pseudo-online analysis of all subjects are shown in Supplementary Fig. 1. Table IV displays the success rates of the pseudo-online analysis across the 15 subjects. An average performance exceeding 0.75 in four sliding windows has been the criterion for

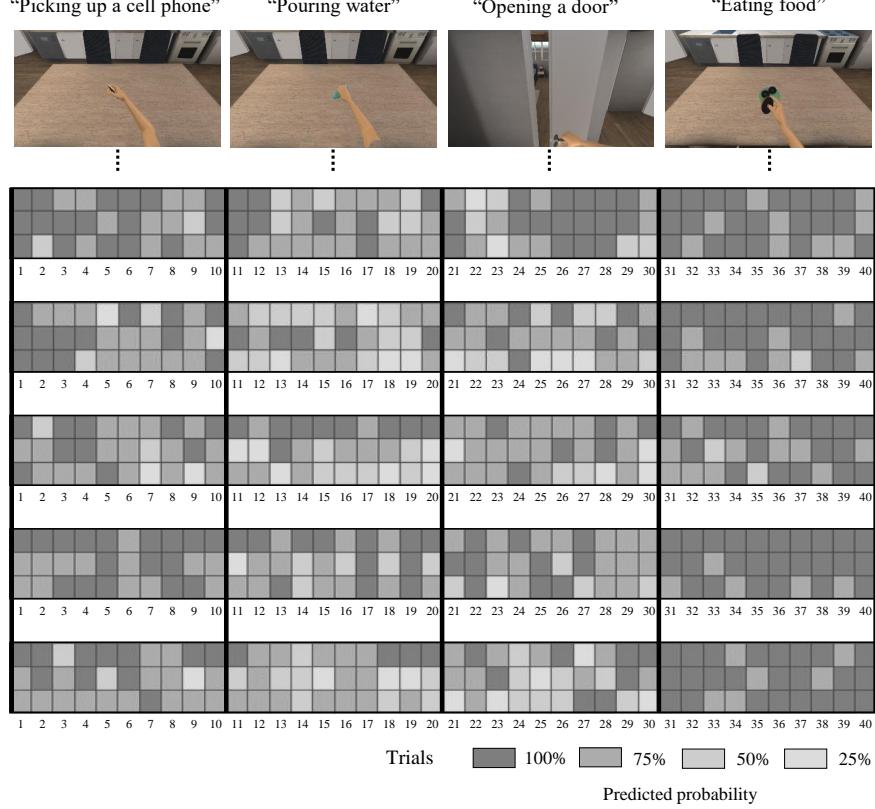


Figure 8: The top three-dimensional virtual environment representations are the visual simulation on the monitor given to the user before the user performed high-level visual imagery. Each of the four classes consists of 10 trials, expressed as follows: trial 1–10; picking up a cell phone (*class-1*), trial 11–20; pouring water (*class-2*), trial 21–30; opening a door (*class-3*), and trial 31–40; eating food (*class-4*). The figure below the visual stimulation is a representation of the classification outputs in the pseudo-online analysis. The darker grey squares indicate higher decoding success rates.

EEG-decoding success. Although we set this criterion strictly, with the proposed network, the average pseudo-online performance of all subjects across the three runs was 0.73. Among all subjects, Sub14 performed the best, scoring 0.91 over 40 trials.

Table 4: Evaluation Performance for Pseudo-Online Analysis through the Success Rate of Decoding

Subjects	Success rate			
	Run I	Run II	Run III	Average
Sub01	0.73 (29/40)	0.83 (33/40)	0.58 (23/40)	0.71 ( $\pm 0.13$ )
Sub02	0.63 (25/40)	0.73 (29/40)	0.63 (25/40)	0.67 ( $\pm 0.05$ )
Sub03	0.73 (29/40)	0.63 (25/40)	0.73 (29/40)	0.63 ( $\pm 0.10$ )
Sub04	0.80 (32/40)	0.70 (28/40)	0.80 (32/40)	0.73 ( $\pm 0.06$ )
Sub05	0.63 (25/40)	0.73 (29/40)	0.63 (25/40)	0.67 ( $\pm 0.05$ )
Sub06	0.70 (28/40)	0.73 (29/40)	0.70 (28/40)	0.71 ( $\pm 0.01$ )
Sub07	0.73 (29/40)	0.48 (19/40)	0.73 (29/40)	0.63 ( $\pm 0.13$ )
Sub08	0.68 (27/40)	0.60 (24/40)	0.68 (27/40)	0.68 ( $\pm 0.09$ )
Sub09	0.90 (36/40)	0.88 (35/40)	0.90 (36/40)	0.88 ( $\pm 0.01$ )
Sub10	0.70 (28/40)	0.63 (25/40)	0.70 (28/40)	0.70 ( $\pm 0.08$ )
Sub11	0.58 (23/40)	0.90 (36/40)	0.63 (25/40)	0.75 ( $\pm 0.14$ )
Sub12	0.60 (24/40)	0.60 (24/40)	0.65 (26/40)	0.64 ( $\pm 0.04$ )
Sub13	0.63 (25/40)	0.78 (31/40)	0.65 (26/40)	0.80 ( $\pm 0.16$ )
Sub14	0.80 (32/40)	0.85 (34/40)	0.88 (35/40)	0.91 ( $\pm 0.08$ )
Sub15	0.73 (29/40)	0.68 (27/40)	0.75 (30/40)	0.78 ( $\pm 0.11$ )

### 3.5. Classification Performance using Leave-One-Subject-Out Cross-Validation Approach

We applied the LOSO approach (Lin et al., 2024; Fazli et al., 2009) to verify the possibility of solving subject-independent issues. In the LOSO method, the network is trained using the data from the source subjects and transferred to test the unknown data from the new subject. In this work, we set the training data to all subjects except one user who became a target. In LOSO conditions, baseline methods and proposed methods were trained using EEG data of 4 subjects, and a total of 16,000 training data were used, considering that the number of training data was 4,000 per subject. We selected the top five subjects who recorded high performance in the subject-dependent condition and confirmed the classification performance in the LOSO condition.

As shown in Table V, the proposed network recorded the highest average classification performance in the LOSO condition compared to the baseline

Table 5: Comparison of Classification Performance under LOSO Condition with Deep Learning Methods

	ConvNet	EEGNet	FBCNet	TSformer	FCDN
Sub04	0.3813 ( $\pm 0.0327$ )	0.3968 ( $\pm 0.0226$ )	0.3845 ( $\pm 0.0237$ )	0.3987 ( $\pm 0.0375$ )	0.3950 ( $\pm 0.0262$ )
Sub06	0.3313 ( $\pm 0.0284$ )	0.3252 ( $\pm 0.0217$ )	0.3472 ( $\pm 0.0195$ )	0.3886 ( $\pm 0.0207$ )	0.4013 ( $\pm 0.0247$ )
Sub09	0.5138 ( $\pm 0.0193$ )	0.5125 ( $\pm 0.0143$ )	0.5320 ( $\pm 0.0229$ )	0.5103 ( $\pm 0.0183$ )	0.5225 ( $\pm 0.0121$ )
Sub14	0.5825 ( $\pm 0.0211$ )	0.6375 ( $\pm 0.0204$ )	0.6067 ( $\pm 0.0253$ )	0.6592 ( $\pm 0.0264$ )	0.6588 ( $\pm 0.0217$ )
Sub15	0.4400 ( $\pm 0.0373$ )	0.4175 ( $\pm 0.0195$ )	0.4287 ( $\pm 0.0194$ )	0.4823 ( $\pm 0.0204$ )	0.5025 ( $\pm 0.0145$ )
Avg.	0.4498 ( $\pm 0.0900$ )	0.4579 ( $\pm 0.2124$ )	0.4598 ( $\pm 0.0845$ )	0.4878 ( $\pm 0.0782$ )	0.4960 ( $\pm 0.0964$ )

methods. In this study, the proposed method, FCDN, demonstrated a notable performance with a classification score of 0.4960, surpassing several established models. Specifically, it outperformed ConvNet (0.4498), EEG-Net (0.4579), FBCNet (0.4598), and TSformer (0.4878). The classification performance of the proposed network in the LOSO condition was similar to that of the subject-dependent condition. These results suggest that the proposed network is robust under subject-independent conditions.

#### 4. Discussion

High-level visual imagery provides a paradigm by which users can effectively imagine an intended movement by only thinking. We identified the neurophysiological features of high-level visual imagery and proposed a network to investigate the feasibility of BCI applications. We have highlighted the possibility of identifying users' intentions in real-time via pseudo-online analysis and demonstrated the possibility of developing a subject-independent system via LOSO.

We have provided a highly intuitive BCI paradigm for subjects to improve EEG signal decoding performance for the development of BCI applications. As a BCI paradigm, we provided high-level visual imagery to subjects in this study, and then performed data analysis to verify the quality of the EEG

data obtained using our paradigm. The measurement of power spectra in the channels responsible for each brain region showed significant results in the prefrontal and occipital regions, indicating a similar tendency to conventional visual imagery-related studies (Sousa et al., 2017; Koizumi et al., 2019; Xie et al., 2020; Llorella et al., 2021; Kilmarx et al., 2024).

We also compared the classification performance of the proposed network with that of conventional decoding approaches. The proposed network recorded a higher performance for the 4-class condition than did the conventional methods, and we deduced that the proposed network is an effective network for decoding high-level visual imagery. We observed a consistent performance trend across subjects for the various deep learning networks, including the proposed model. However, this trend was not aligned with the performance pattern observed in the FBCSP approach, indicating that the subject-wise classification tendencies differ significantly between FBCSP and deep learning-based methods. For visual imagery, both spatial and temporal information is important. Consequently, the FBCSP approach, which does not emphasize temporal information, demonstrated lower performance and the constant tendency among subjects was eliminated.

As high-level visual imagery data are based on the user's imagined temporal changes in a 3D space, we designed the proposed network accordingly. The network proposed in this study emphasized channels containing meaningful features related to high-level visual imagery based on PLVs and used these to modify the raw EEG data. In addition, the channel-wise separable convolutional layer was placed on the last layer of the CNN. Features were extracted using the two preceding convolutional layers to include temporal information, and these features were used as inputs for DeiT. This structure extracted information based on features, including temporal information, when extracting the spatial features of the input using DeiT. Thus, the output contains both spatial and temporal information.

In the pseudo-online analysis, the final decisions were derived using four segmented windows, which yielded a high decoding performance. This confirms that high-level visual imagery can be beneficial in real-life applications for users, and demonstrates the effectiveness of the proposed network in decoding high-level visual imagery. Notably, this technology shows potential in aiding individuals through BCI-based applications, such as robotic arms, suggesting a promising avenue for practical implementation. While this study was conducted with a relatively small sample size of 15 participants, consistent decoding patterns were observed across subjects. The robust classi-

Table 6: The Pearson Correlation Coefficient Results for Phase-Locking Value in Representative Subjects

	r-vlaue				
	Sub04	Sub06	Sub09	Sub14	Sub15
Sub04	1.0000	0.7881	0.8555	0.7390	0.4025
Sub06	-	1.0000	0.5627	0.6523	0.4864
Sub09	-	-	1.0000	0.6612	0.2755
Sub14	-	-	-	1.0000	0.4495
Sub15	-	-	-	-	1.0000

fication performance under the LOSO validation scheme suggests that the proposed method captures generalizable neural features across individuals.

For all subjects, all deep learning approaches yielded a classification performance of  $\geq 0.25$ , which is the chance level. The proposed network yielded the highest level of performance. The high-level visual imagery is possible in a subject-independent approach because it demonstrated a constant data distribution, regardless of the subject, and the spatial area emphasized in the network was similar.

As shown in Table VI, the similarity of PLVs among the subjects was calculated using Pearson’s correlation coefficient (CC) (Benesty et al., 2009)  $r$ , which refers to the similarity of PLVs among subjects, and were mostly above 0.3, with a distinct amount of correlation. The CCs of all subjects are shown in Supplementary Table III. As most users have PLVs with distinct correlations, we could infer that users were imagining in a similar manner, which could be the starting point for the development of a subject-independent approach. All three cases ( $r \leq 0.3$ ) were related to Sub15, indicating that Sub15 used a different pattern of imagination from other subjects. If re-learning is requested from these types of subjects or a deep learning approach that minimizes spatial information is applied, better performance could be expected. In this study, we constructed a network in which spatial information was emphasized based on PLVs, which can be used as an indicator of data learning in subject-independent conditions.

## 5. Conclusions and Future Works

In this study, we applied a 3D-BCI training platform allowing users to perform high-level visual imagery effectively. Using the 3D-BCI training

platform, we were able to collect high-quality visual imagery data from participants, and we verified the quality of the data using neurophysiological methods. In addition, we proposed FCDN to decode the obtained visual imagery data. The network was designed to consider both spatial and temporal information based on the characteristics of visual imagery. Using functional connectivity, we emphasized channels that were highly correlated with visual imagery, and accordingly extracted spatial information using convolutional layers. Subsequently, we used the extracted features as inputs to DeiT. The network proposed in this study recorded a robust decoding performance regardless of the number of classes. Achieving high classification performance for various classes is important, but it is also significant that the types of classes proposed in this study were designed to involve complex upper-limb movements. While the current study involved 15 participants, the consistent subject-wise decoding performance suggests that the proposed method captures generalizable neural representations. Nevertheless, we recognize the importance of further validation with a broader population. To this end, we plan to conduct additional experiments to collect data from more participants, aiming to enhance the generalizability and robustness of the proposed method. We will also further develop FCDN, the proposed network, for high classification performance, so that it can be applied to real-time BCI scenarios. In addition, we will explore methods to reduce the computational complexity of the networks proposed in this study. We also aim to expand the dataset by increasing the number of movement classes, allowing the model to decode a broader range of upper-limb movements. Based on this, we will apply the proposed network to an EEG-based robotic arm system, which will contribute to the development of rehabilitation systems for patients with tetraplegia or for supporting the daily life activities of healthy people.

## Acknowledgements

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University)) and the National Research Foundation of Korea (NRF) grant funded by the MSIT (No.2022-2-00975, MetaSkin: Developing Next-generation Neurohaptic Interface Technology that enables Communication and Control in Metaverse by Skin Touch.

## References

- Ahn, H.J., Lee, D.H., Jeong, J.H., Lee, S.W., 2022. Multiscale convolutional transformer for EEG classification of mental imagery in different modalities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31, 646–656.
- Ai, J., Meng, J., Mai, X., Zhu, X., 2023. Bci control of a robotic arm based on ssvep with moving stimuli for reach and grasp tasks. *IEEE Journal of Biomedical and Health Informatics* 27, 3818–3829.
- Amin, S.U., Alsulaiman, M., Muhammad, G., Mekhtiche, M.A., Hossain, M.S., 2019. Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion. *Future Generation Computer Systems* 101, 542–554.
- Ang, K.K., Chin, Z.Y., Zhang, H., Guan, C., 2008. Filter bank common spatial pattern (FBCSP) in brain-computer interface. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* , 2390–2397.
- Arsalan, A., Majid, M., Butt, A.R., Anwar, S.M., 2019. Classification of perceived mental stress using a commercially available EEG headband. *IEEE Journal of Biomedical and Health Informatics* 23, 2257–2264.
- Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. Noise reduction in speech processing , 1–4.
- Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L.E., Sturm, I., Curio, G., Müller, K.R., 2010. The Berlin brain-computer interface: non-medical uses of BCI technology. *Frontiers in Neuroscience* 4, 198.
- C. I. Penalosa and S. Nishio, 2018. BMI Control of a third arm for multi-tasking. *Science Robotics* 3, 1–6.
- Carrier, J., Land, S., Buysse, D.J., Kupfer, D.J., Monk, T.H., 2001. The effects of age and gender on sleep EEG power spectral density in the middle years of life (ages 20–60 years old). *Psychophysiology* 38, 232–242.

- Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 .
- Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134, 9–21.
- Dong, C., Loy, C.C., He, K., Tang, X., 2015. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 295–307.
- Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.R., Grozea, C., 2009. Subject-independent mental state classification in single trials. Neural Networks 22, 1305–1312.
- Han, C.H., Müller, K.R., Hwang, H.J., 2020. Enhanced performance of a brain switch by simultaneous use of EEG and NIRS data for asynchronous brain-computer interface. IEEE Transactions on Neural Systems and Rehabilitation Engineering 28, 2102–2112.
- Honey, C.J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.P., Meuli, R., Hagmann, P., 2009. Predicting human resting-state functional connectivity from structural connectivity. Proceedings of the National Academy of Sciences of the United States of America 106, 2035–2040.
- Jebelli, H., Khalili, M.M., Lee, S., 2018. A continuously updated, computationally efficient stress recognition framework using electroencephalogram (EEG) by applying online multitask learning algorithms (OMTL). IEEE Journal of Biomedical and Health Informatics 23, 1928–1939.
- Jeong, J.H., Kim, K.T., Kim, D.J., Lee, S.J., Kim, H., 2022. Subject-transfer decoding using the convolutional neural network for motor imagery-based brain-computer interface, in: International Conference Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 48–51.
- Jeong, J.H., Shim, K.H., Kim, D.J., Lee, S.W., 2020. Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals. IEEE Transactions on Neural Systems and Rehabilitation Engineering 28, 1226–1238.

- Kilmarx, J., Tashev, I., Millán, J.d.R., Sulzer, J., Lewis-Peacock, J., 2024. Evaluating the feasibility of visual imagery for an EEG-based brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* .
- Koizumi, K., Ueda, K., Tateyama, N., Nakao, M., 2019. EEG source analysis of visual motion imagery for application to brain-computer interface, in: GBCIC.
- Kwak, N.S., Lee, S.W., 2019. Error correction regression framework for enhancing the decoding accuracies of ear-EEG brain-computer interfaces. *IEEE Transactions on Cybernetics* 50, 3654–3667.
- Kwon, B.H., Jeong, J.H., Cho, J.H., Lee, S.W., 2020. Decoding of intuitive visual motion imagery using convolutional neural network under 3D-BCI training environment, in: International Conference Proceedings of the IEEE Systems, Man, and Cybernetics Society (SMC), pp. 2966–2971.
- Lachaux, J.P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. *Human Brain Mapping* 8, 194–208.
- Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering* 15, 1–17.
- Lin, L., Li, P., Wang, Q., Bai, B., Cui, R., Yu, Z., Gao, D., Zhang, Y., 2024. An EEG-based cross-subject interpretable cnn for game player expertise level classification. *Expert Systems with Applications* 237, 121658.
- Liu, W., Guo, C., Gao, C., 2024. A cross-session motor imagery classification method based on Riemannian geometry and deep domain adaptation. *Expert Systems with Applications* 237, 121612.
- Llorella, F.R., Iáñez, E., Azorín, J.M., Patow, G., 2021. Classification of imagined geometric shapes using EEG signals and convolutional neural networks. *Neuroscience Informatics* 1, 100029.
- Luo, T.j., 2024. Selective multi-view time-frequency decomposed spatial feature matrix for motor imagery EEG classification. *Expert Systems with Applications* 247, 123239.

- Makeig, S., 1993. Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones. *Electroencephalography and Clinical Neurophysiology* 86, 283–293.
- Mane, R., Chew, E., Chua, K., Ang, K.K., Robinson, N., Vinod, A.P., Lee, S.W., Guan, C., 2021. Fbcnet: A multi-view convolutional neural network for brain-computer interface. arXiv preprint arXiv:2104.01233 .
- Maris, E., Oostenveld, R., Aug. 2007. Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods* 164, 177–190.
- Mulkey, M.A., Huang, H., Albanese, T., Kim, S., Yang, B., 2023. Supervised deep learning with vision transformer predicts delirium using limited lead eeg. *Scientific Reports* 13, 7890.
- Pearson, J., 2019. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience* 20, 624–634.
- Pérez-Velasco, S., Santamaría-Vázquez, E., Martínez-Cagigal, V., Marcos-Martínez, D., Hornero, R., 2022. EEGSym: Overcoming inter-subject variability in motor imagery based BCIs with deep learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30, 1766–1775.
- Roach, B.J., Mathalon, D.H., 2008. Event-related EEG time-frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia. *Schizophrenia Bulletin* 34, 907–926.
- Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T., 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping* 38, 5391–5420.
- Sousa, T., Amaral, C., Andrade, J., Pires, G., Nunes, U.J., Castelo-Branco, M., 2017. Pure visual imagery as a potential approach to achieve three classes of control for implementation of BCI in non-motor disorders. *Journal of Neural Engineering* 14, 046026.
- Suk, H.I., Lee, S.W., 2012. A novel bayesian framework for discriminative feature extraction in brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 286–299.

- Sun, L., Feng, Z., Lu, N., Wang, B., Zhang, W., 2019. An advanced bispectrum features for EEG-based motor imagery classification. *Expert Systems with Applications* 131, 9–19.
- Sun, Y., Li, Y., Chen, Y., Yang, C., Sun, J., Liang, L., Chen, X., Gao, X., 2024. Efficient dual-frequency SSVEP brain-computer interface system exploiting interocular visual resource disparities. *Expert Systems with Applications* 252, 124144.
- Tibrewal, N., Leeuwis, N., Alimardani, M., 2022. Classification of motor imagery EEG using deep learning increases performance in inefficient BCI users. *PloS One* 17, e0268880.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention, in: International conference on machine learning, PMLR. pp. 10347–10357.
- Wang, B., Wong, C.M., Kang, Z., Liu, F., Shui, C., Wan, F., Chen, C.P., 2020. Common spatial pattern reformulated for regularizations in brain–computer interfaces. *IEEE Transactions on Cybernetics* 51, 5008–5020.
- Wang, Y., Wang, Y.T., Jung, T.P., 2012. Translation of EEG spatial filters from resting to motor imagery using independent component analysis. *PloS One* 7, e37665.
- Xie, S., Kaiser, D., Cichy, R.M., 2020. Visual imagery and perception share neural representations in the alpha frequency band. *Current Biology* 30, 2621–2627.
- Zhang, D., Yao, L., Chen, K., Wang, S., Chang, X., Liu, Y., 2019. Making sense of spatio-temporal preserving representations for EEG-based human intention recognition. *IEEE Transactions on Cybernetics* 50, 3033–3044.
- Zhang, T., Zheng, W., Cui, Z., Zong, Y., Li, Y., 2018. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Transactions on Cybernetics* 49, 839–847.
- Zhou, Y., Yu, T., Gao, W., Huang, W., Lu, Z., Huang, Q., Li, Y., 2023. Shared three-dimensional robotic arm control based on asynchronous BCI

and computer vision. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31, 3163–3175.