FGANet: fNIRS-Guided Attention Network for Hybrid EEG-fNIRS Brain-Computer Interfaces

Youngchul Kwak, Woo-Jin Song, *Life Member, IEEE*, and Seong-Eun Kim, *Member, IEEE*

***Abstract*—Non-invasive brain-computer interfaces (BCIs) have been widely used for neural decoding, linking neural signals to control devices. Hybrid BCI systems using electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) have received significant attention for overcoming the limitations of EEG- and fNIRS-standalone BCI systems. However, most hybrid EEG-fNIRS BCI studies have focused on late fusion because of discrepancies in their temporal resolutions and recording locations. Despite the enhanced performance of hybrid BCIs, late fusion methods have difficulty in extracting correlated features in both EEG and fNIRS signals. Therefore, in this study, we proposed a deep learning-based early fusion structure, which combines two signals before the fully-connected layer, called the fNIRS-guided attention network (FGANet). First, 1D EEG and fNIRS signals were converted into 3D EEG and fNIRS tensors to spatially align EEG and fNIRS signals at the same time point. The proposed fNIRS-guided attention layer extracted a joint representation of EEG and fNIRS tensors based on neurovascular coupling, in which the spatially important regions were identified from fNIRS signals, and detailed neural patterns were extracted from EEG signals. Finally, the final prediction was obtained by weighting the sum of the prediction scores of the EEG and fNIRS-guided attention features to alleviate performance degradation owing to delayed fNIRS response. In the experimental results, the FGANet significantly outperformed the EEG-standalone network. Furthermore, the FGANet has 4.0% and 2.7% higher accuracy than the state-of-the-art algorithms in mental arithmetic and motor imagery tasks, respectively.**

***Index Terms*—Brain-computer interface (BCI), deep learning, electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), hybrid BCI, and fNIRS-guided attention networks.**

I. Introduction

THE brain-computer interfaces (BCIs) provide a direct interface between neural activities containing the user's intention and control signals for external devices [1]–[3]. The BCI system may allow individuals to operate assistive devices, such as robot arms [4], wheelchairs [5], and spelling [6]. Furthermore, BCI systems can be used to detect neurological diseases, such as seizures and Alzheimer's disease [7]–[9].

BCI approaches can be divided into invasive and non-invasive BCIs. As invasive BCIs usually record neural signals from electrodes implanted into the brain, non-invasive BCIs are preferred for humans because of their safety and convenience. In non-invasive BCI systems, various neural activities have been utilized, such as electroencephalography (EEG) [5], [6], [9], [10], magnetoencephalography (MEG) [11], [12], functional near-infrared spectroscopy (fNIRS) [13]–[16], and functional magnetic resonance imaging (fMRI) [17], [18]. Despite MEG and fMRI having excellent spatial resolution to study the underlying neuronal activities and cerebral blood flow changes, they are inappropriate for real-world BCI systems because of their large size and high cost. However, EEG and fNIRS are more suitable for real-world applications because of their portability and low cost. Therefore, EEG- or fNIRS-based BCI systems have been extensively investigated.

EEG and fNIRS measure different physiological dynamics of brain activity. EEG can capture the macroscopic temporal dynamics of neuronal electrical activity through multi-channel electrodes on the scalp. In particular, it has the superior advantage of high temporal resolution with a fast response to stimuli, thus it is popular in medical and engineering applications. However, EEG is vulnerable to movement artifacts and electrical noise. Hence, EEG-standalone BCI systems often misclassify resting-state EEG signals as commands while the subject is not performing any tasks [19]. Compared with EEG, fNIRS is a scalp-based optical spectroscopic measurement that uses a light injection source and detection to measure hemodynamic fluctuations caused by brain activity. Increased neural activity results in increased oxygen consumption to fulfill the demand of the neuronal tissues, which causes a decrease in oxygenated hemoglobin (HbO) and an increase in deoxygenated hemoglobin (HbR). fNIRS is robust to motion artifacts and electrical noise, but has significantly poor temporal resolution and delayed hemodynamic response, making it challenging to construct real-time BCI applications. The maximum classification accuracy of the fNIRS-standalone BCI system was delayed up to 7 s compared to that of the EEG-standalone BCI system for the same task [10]. Therefore, hybrid EEG-fNIRS BCI systems have been introduced to

Manuscript received October 25, 2021; revised January 13, 2022; accepted February 4, 2022. Date of publication February 7, 2022; date of current version February 16, 2022. This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) under Grant 2020M3C1B8081320 and Grant 2020R1C1C1011857. *(Corresponding author: Seong-Eun Kim.)*

Youngchul Kwak and Woo-Jin Song are with the Department of Electronics Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Gyeongbuk 37673, South Korea (e-mail: kyc2058@postech.ac.kr; wjsong@postech.ac.kr).

Seong-Eun Kim is with the Department of Applied Artificial Intelligence, Seoul National University of Science and Technology (SeoulTech), Nowon-gu, Seoul 01811, South Korea (e-mail: sekim@seoultech.ac.kr).

Digital Object Identifier 10.1109/TNSRE.2022.3149899

overcome the limitations of EEG- or fNIRS-standalone BCI systems.

Hybrid EEG-fNIRS BCI systems can significantly improve EEG or fNIRS-standalone BCI systems by combining the advantages of each signal. However, multimodal fusion is challenging because two signals are significantly different in their temporal resolution and recording locations, which can make the joint representation of two signals difficult. Therefore, most of the traditional studies focused on feature- or decision-level fusion techniques, which can simply improve performance by combining hand-crafted EEG and fNIRS features or decision scores [20]–[24]. These methods extract each EEG and fNIRS feature individually, and then the concatenated features are used for classification by linear discrimination analysis (LDA) or support vector machine (SVM). For example, Shin *et al.* [20] extracted prediction scores from the EEG and fNIRS signals, respectively, and used an LDA-based meta-classifier to obtain the final prediction score. Jiang *et al.* [25] proposed an independent decision path fusion (IDPF) method that extracts an independent decision score from each EEG and fNIRS features, including the power spectrum of EEG signals and mean value of the HbO and HbR, and then fused EEG and fNIRS features according to the score to classify brain signals. They developed reliable-based decision-level fusion, which assigns different weights to the decision scores based on their respective accuracy. They developed reliable-based decision-level fusion, which give different weights to decision score based on their respective accuracy. In contrast to the feature-level fusion method, some traditional studies have utilized fNIRS signals as a supplementary tool in EEG-based BCI systems. For example, fNIRS signals are used as predictors of EEG activity to improve the stability of the BCI system [10], [26]. In another study, fNIRS signals were used to find region-specific information related to the task and to apply spatial attention to those regions [27].

In recent years, deep learning techniques have evolved, and they have shown good performance in various research areas such as speech recognition, image classification, and video recognition [28]–[31]. In brain decoding, the deep learning-based approaches have also attracted significant attention in several unimodal (EEG or fNIRS) and hybrid EEG-fNIRS BCI systems [13], [14], [32]–[36] because of the ability to extract high-level representations and classify them directly from a dataset. In deep learning-based hybrid EEG-fNIRS BCI systems, most deep learning structures are designed based on the late fusion method where two signals are merged after fully-connected layer. The main difference between traditional fusion approaches and deep learning approaches is the manner in which unimodal or fusion features are extracted. The deep learning-based approaches extract high-level representation from a convolutional and fully-connected layer. For example, Chiarelli *et al.* [32] concatenated EEG and fNIRS features, and then fed them into an artificial neural network to extract high-level feature representations. Sun *et al.* [33] extracted the deep features of each EEG and fNIRS signal, and then fused them using the pth order polynomial fusion (pth-PF) algorithm with tensor decomposition to deal with an unacceptably large number of parameters. However, late fusion techniques cannot fully capture the underlying homogeneity in a mixed feature space [37]. In particular, Neverova *et al.* [38] highlighted the huge benefits of early fusion in similar modalities, such as red-green-blue (RGB) & depth images [39], radar & infrared sensors [40], and optical flow images & video [41]. EEG and fNIRS are also closely correlated with spatial manners because of neurovascular coupling, thus the reason for neural activities to cause a subsequent change in cerebral blood flow [42]–[44]. Nevertheless, early fusion methods have not been extensively studied in deep-learning-based hybrid EEG-fNIRS BCI systems.

In the EEG-fNIRS fusion structure, it is crucial to extract a joint representation for the underlying physiological spatial correlation between two signals to maximize the performance of hybrid EEG-fNIRS BCI systems. Therefore, we propose a deep learning-based fusion method, known as the fNIRS-guided attention network (FGANet). To enable the use of spatial information over the scalp, one-dimensional (1D) EEG and fNIRS signals are projected into a two-dimensional space, and the temporal information is assigned for each corresponding point. Thereafter, the joint representation is extracted from the fNIRS-guided attention (FGA) layer designed based on neurovascular coupling. This layer extracts detailed information of neural activities from EEG signals, while the fNIRS signals are only guided to the spatially important region if two signals can be simultaneously obtained over the whole scalp in a hybrid EEG-fNIRS BCI system. This is because the EEG signal can capture temporal dynamic neural activities in a short time window, while fNIRS signals can obtain the area where neurons are activated by stimuli with low electrical noise and movement artifacts in the hybrid EEG-fNIRS BCI system. However, due to the delayed fNIRS response, the performance of the FGA layer may deteriorate at the beginning of the trial. As a result, we proposed a prediction method to mitigate the inconsistency in spatial neurovascular coupling caused by the delayed response, wherein the prediction scores of EEG and fusion branches are considered together for the final classification. If the reliability of the prediction score of the fusion branch is low because of the mismatch between the EEG and fNIRS responses, it is better to assign more weight to the prediction score of the EEG branch. Subsequently, the prediction weight is determined adaptively depending on the importance of the EEG and fusion features in the learning process for a higher classification accuracy. The final prediction is achieved by the weighted sum of the prediction scores of the EEG and the fusion branches.

The experimental results showed that fNIRS guided a spatially important region for brain decoding. In addition, the prediction weight on the fusion feature increases as time passes on the stimuli, and vice versa for the EEG feature. Furthermore, our proposed fusion model significantly enhanced the performance of the EEG- and fNIRS-standalone BCI systems. In summary, the main contributions of FGANet are as follows.

1) We proposed the spatially aligned method for EEG and fNIRS signals by converting 1D EEG and fNIRS signals into 3D EEG and fNIRS tensors.

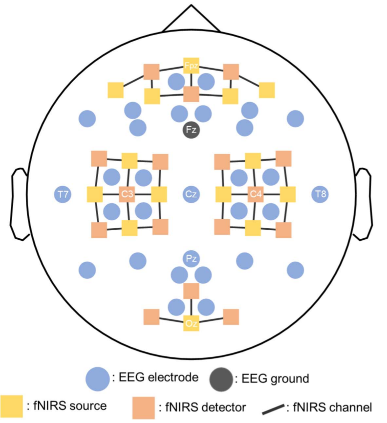


Fig. 1. The placement of EEG electrodes, fNIRS sources and fNIRS detectors.

2) An early fusion method, named the fNIRS-guided attention layer, was proposed, where fNIRS guides the important region for brain decoding and applies spatial attention to EEG features.
3) A prediction method was developed to alleviate the deterioration of decoding performance caused by the inherent delay of fNIRS signals.

The remainder of this paper is organized as follows. Section II describes the dataset used to evaluate the proposed algorithm. In Section III, we propose the EEG- and fNIRS-standalone deep neural network structure and a framework of FGANet. The experimental results and analysis are discussed in section IV. Finally, Section V presents the conclusions.

II. Dataset

A public dataset [20], which simultaneously records EEG and fNIRS, was utilized in this study. The data were acquired from twenty-eight right-handed subjects and one left-handed subject (14 males and 15 females) with an average age of 28.5 ± 3.7 years (mean ± standard deviation). EEG signals were recorded at 1000 Hz from 30-channels (Fig. 1), and fNIRS signals were recorded at 12.5 Hz from thirty-six channels consisting of 14 sources and 16 detectors (Fig. 1). Thereafter, the EEG and fNIRS signals were downsampled to 200 Hz and 10 Hz, respectively, by the data provider.

Subjects were required to perform 30 trials for each task: baseline (BS) as a rest state condition, mental arithmetic (MA), left-hand motor imagery (MI), and right-hand MI. The subject rested without any thoughts during the BS condition. In the MA task, the subject was required to repeatedly subtract the one-digit number from the three-digit number (e.g., 384 – 8) during the task period. For the MI task, subjects conducted kinesthetic MI to imagine the opening and closing of their hands. The trial started with 2 s of a visual introduction of the task, followed by 10 s of a task period and resting period, which was provided randomly from 15 to 17 s (Fig. 2). A detailed data description is provided in [20].

The EEG signals were re-referenced with a common average reference, filtered at 0.5—50Hz and then downsampled to 120 Hz. Thereafter, electrooculography (EOG) artifacts were removed by independent component analysis (ICA).

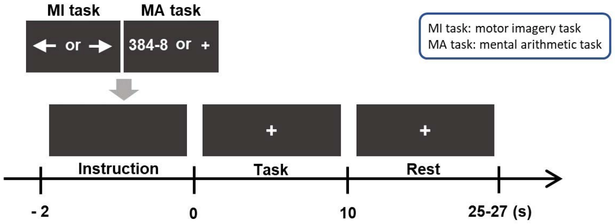


Fig. 2. The paradigm of the experimental process.

We divided the data into two datasets according to the task as done in [20], [33]: MA dataset (baseline vs. MA) and MI dataset (left-hand MI vs. right-hand MI). Instead of using a 10 s task period to train the network, we cropped 3 s both EEG and fNIRS signals with a time step of 1 s to evaluate real-time BCI performance, as done in [20], [33]. Therefore, the size of the EEG and fNIRS signals were 30 × 360 (channel × time) and 36 × 30 (channel × time), respectively, with a total of 1,740 (29 subject × 30 trials × 2 task) trials for each dataset.

III. Method

A. 3D Tensor Generation

1) 3D EEG Tensor: The spatiotemporal dynamics of the brain represent a complex cognitive process. For example, theta oscillations (4-8 Hz) in the frontal cortex are related to cognitive workload [45]. In addition, alpha oscillations (8-12 Hz) of the parietal cortex represent visual attention [46], and beta oscillations (15-25 Hz) of sensorimotor regions are correlated with the mental simulation of actions [47]. Therefore, to include spatiotemporal information in the input data for training the network, we converted 1D EEG signals to 3D EEG tensors. To obtain 3D EEG images, we projected the 3D electrode locations on the scalp into a 2D image with size 16 × 16 using azimuthal equidistant projection as in [34] (Fig. 3(a)). The data of the mapped point were filled with temporal information of the corresponding electrode. Thereafter, the empty values between the electrodes were interpolated using cubic spline interpolation. Accordingly, the generated 3D EEG images $\boldsymbol{X}^{\text{eeg}} \in \mathbb{R}^{16\times16\times360}$ contained spatial information in the first two dimensions and temporal information in the last dimension.

2) 3D fNIRS Tensor: fNIRS consists of a source emitting near-infrared light and a detector receiving light that diffuses out of the brain tissue. To obtain hemodynamic changes by neural activity, detected raw fNIRS signals should be converted into changes in HbO and HbR. Because HbO and HbR have different absorption coefficients for different wavelengths of near-infrared light, the change in HbO and HbR can be obtained by the ratio of incident light intensity to the detected light intensity for the two different wavelengths of light. We employed the Beer-Lambert equation [48] to obtain the concentration change of HbO and HbR in the task period relative to those of the baseline time window. The baseline

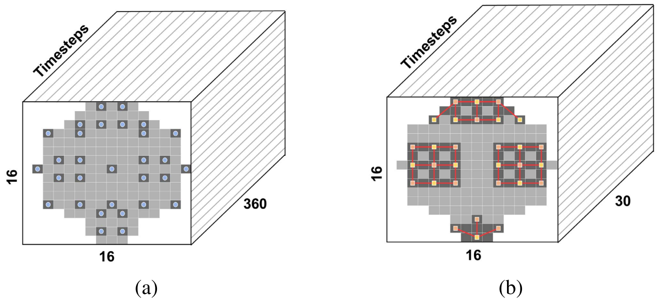


Fig. 3. The visualization of the (a) 3D EEG tensor and the (b) 3D fNIRS tensor.

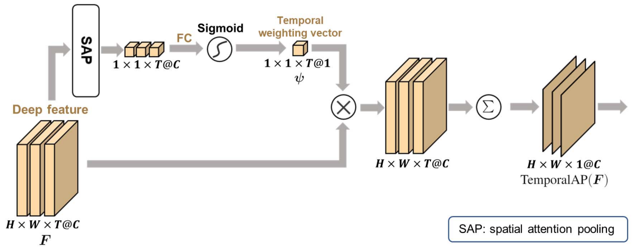


Fig. 4. The architecture of the temporal attentive pooling layer.

TABLE I
THE STRUCTURE OF THE ESNET

Layer	Kernal			Output
	Filter size	Stride	Channel	Dimension
Input				16x16x360x1
Conv1	2x2x9	2x2x4	16	8x8x90x16
Conv2	2x2x3	2x2x2	32	4x4x45x32
Conv3	2x2x3	2x2x2	64	2x2x25x64
TemporalAP				2x2x1x64
FC	256x64			1x64
Dropout				1x64
Softmax	64x2			2

interval is defined as the time windows from −5 to −2 s at the start of the trial.

EEG and fNIRS signals are closely correlated with spatial manners because of neurovascular coupling [42]–[44]. Therefore, to utilize the spatial correlation between two signals, we converted the 1D fNIRS signal into a 3D fNIRS tensor using a similar 3D EEG tensor generation process. Contrary to EEG signals, fNIRS signals are measured by light source and detector. Therefore, the path between the source and detector was filled with the same HbO or HbR value, as shown in Fig. 3(b). Thereafter, the empty values between source and detector were interpolated using spline interpolation, as conducted in EEG signals. Consequently, the 3D fNIRS image $\boldsymbol{X}^{\text{fnirs}} \in \mathbb{R}^{16\times16\times30}$ was spatially aligned with the 3D EEG image $\boldsymbol{X}^{\text{eeg}}$.

B. 3D CNN Structure for Unimodal Signal

In this subsection, we describe the structure of the EEG-standalone network (ESNet) and fNIRS-standalone network (FSNet), which are summarized in Tables I and II.

TABLE II
THE ARCHITECTURE OF THE FSNET

Layer	Kernal			Output
	Filter size	Stride	Channel	Dimension
Input				16x16x60x1
Conv1	2x2x9	2x2x2	16	8x8x30x16
Conv2	2x2x3	2x2x2	32	4x4x15x32
Conv3	2x2x3	2x2x2	64	2x2x8x64
TemporalAP				2x2x1x64
FC	256x64			1x64
Dropout				1x64
Softmax	64x2			2

Because 3D convolution can extract both spatial and temporal information, 3D CNNs achieve superior performance in various areas where input data contain 3D data, such as video recognition and EEG decoding [34], [35], [49], [50]. Therefore, we constructed ESNet and FSNet consisting of three 3D convolutional layers with a rectified linear unit (ReLU) activation function to extract spatiotemporal information from 3D EEG and fNIRS tensors, respectively. The deep features were downsampled by the stride of the convolutional layer. To align the spatial dimension between EEG and fNIRS features, we set identical filter sizes and strides for the spatial dimension of the two networks, but different filter sizes and strides were designed for the temporal dimension because the temporal resolutions are different between the two signals. A small kernel size (2 × 2 × 3) and stride (2 × 2 × 2) are used to increase the non-linearity with fewer parameters following the paper [29]. However, because the temporal dimension is much larger than the spatial dimension, we use a larger kernel size (2 × 2 × 9) and stride (2 × 2 × 4) in the first convolutional layer for the temporal dimension, and these values are determined through the optimization process described in result section. To efficiently compress the temporal information of the 3D feature extracted by three consecutive convolutional layers, we developed a temporal attentive pooling layer. Finally, the compressed feature was classified by one fully connected (FC) layer with the ReLU function and softmax layer. Dropout [51] was applied to the output of the FC layer to prevent overfitting.

The 3D deep feature includes $\boldsymbol{F} \in \mathbb{R}^{H\times W\times T\times C}$, where H and W are the height and width of the spatial dimension, respectively, T is the total number of time steps, and C is the number of channels containing temporal information for both EEG and fNIRS signals. In particular, brain activity considerably changes with time while performing a task, and thus brain signals may have task-related segments at a specific time index over the task period. For instance, the difference in spectral power owing to cognitive load was presented at 1-2 s after the stimuli, and it varies according to the subject [52]. Therefore, it is important to provide a large weight to task-related time segments. Consequently, we proposed the temporal attentive pooling (TAP) layer, TAP : $\mathbb{R}^{H\times W\times T\times C} \rightarrow \mathbb{R}^{H\times W\times 1\times C}$, defined as

$$\text{TAP}(\boldsymbol{F})_{h,w,1,c} = \sum_{t'} \psi_{t'} F_{h,w,t',c}, \tag{1}$$

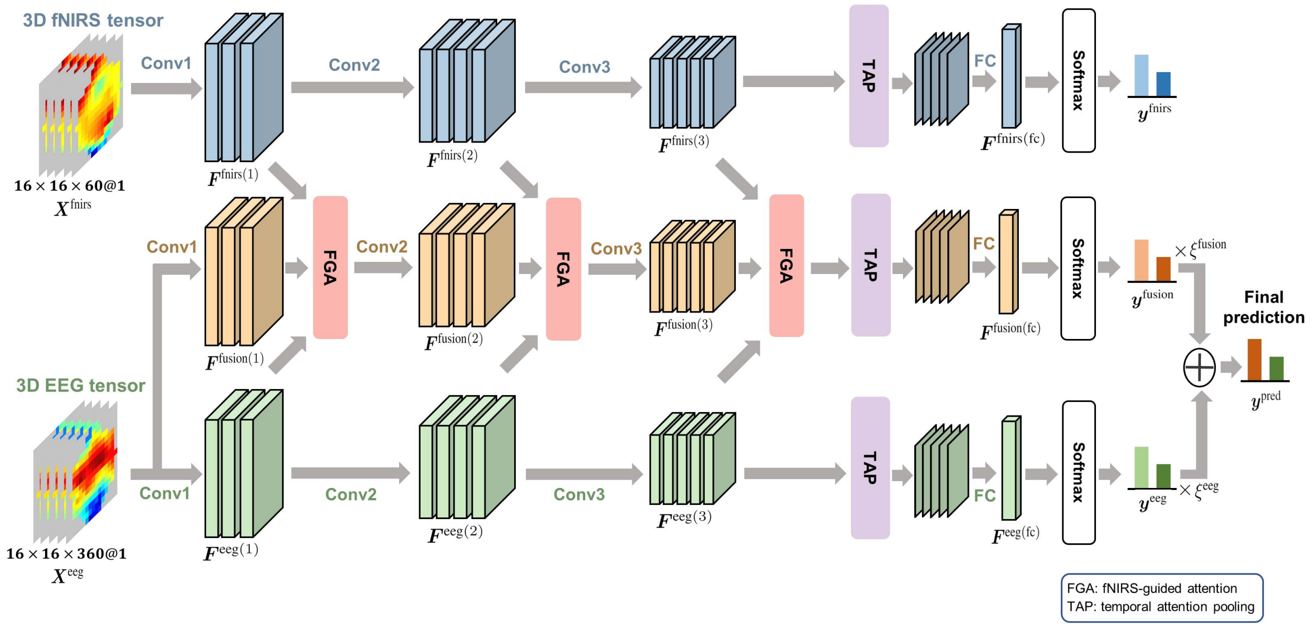


Fig. 5. The architecture of the FGANet.

where $F_{h,w,t,c}$ represents the (h, w, t, c) component of the deep feature $\boldsymbol{F}$, and $\boldsymbol{\psi} \in \mathbb{R}^{T\times 1}$ is a temporal weight vector that is adaptively determined by the input signals. To obtain the temporal weight vector, the 3D deep feature $\boldsymbol{F}$ was spatially reduced by spatial attentive pooling (SAP), and then fed into the FC layer with a softmax function as follows:

$$\boldsymbol{\psi} = \sigma_{\text{soft}}\Big(\text{FC}^{T\times 1}\big(\text{SAP}(\boldsymbol{F})\big)\Big), \tag{2}$$

where

$$\text{SAP}(\boldsymbol{F})_{1,1,t,c} = \sum_{h',w'} \Theta_{h',w'} F_{h',w',t,c}, \tag{3}$$

$\sigma_{\text{soft}}(\cdot)$ represents the softmax function, $\text{FC}^{T\times 1}(\cdot)$ is the FC layer with an output size of $T \times 1$, and $\boldsymbol{\Theta} \in \mathbb{R}^{H\times W}$ is a learnable parameter for spatial attentive pooling. Spatial attentive pooling compresses deep features with a spatial domain by weighting task-related brain regions. The process of TAP is illustrated in Fig. 4.

C. fNIRS-Guided Attention Network (FGANet)

Here, we propose an fNIRS-guided attention network (FGANet) to achieve a high-performance hybrid EEG-fNIRS BCI system. The FGANet consists of three feature extractor branches: the EEG branch, fNIRS branch, and fusion branch, as shown in Fig. 5. We utilized the same three convolutional layers of ESNet and FSNet as feature extractors of the EEG and fNIRS branches, respectively. The fusion branch has an fNIRS-guided attention (FGA) layer after each of the three convolutional layers that are equivalent to those of the ESNet. The FGA layer is designed to extract the joint representation of the 3D EEG and fNIRS tensors based on neurovascular coupling, thus the fNIRS and EEG signals are strongly correlated with the spatial dimension because active brain activity in a specific brain region promotes both cortical currents and blood flow [42]–[44].

EEG can capture the temporal dynamics of neural activities in a short time, but various artifacts are apt to contaminate it. However, fNIRS has a low temporal resolution compared to EEG, but is robust to electrical noise and motion artifacts. To acquire the good points of both brain signals, we proposed an FGA layer, where fNIRS is used to extract spatially important regions for brain decoding, and spatial attention is applied to EEG features. The proposed FGA is different from the conventional self-attention mechanism, which emphasizes the essential region for classification by multiplying the original feature and attention map obtained from the original feature itself. In our work, we proposed an FGA layer based on the characteristics of brain signals, where fNIRS extracts spatially important regions for brain decoding and applies spatial attention to the fusion feature.

The output of the FGA layer $\hat{\boldsymbol{F}}^{\text{fusion}}$ is defined as

$$\hat{F}^{\text{fusion}}_{h,w,t,c} = \gamma\, F^{\text{eeg}}_{h,w,t,c} + (1-\gamma) F^{\text{fusion}}_{h,w,t,c} + \Phi_{h,w,1,1} F^{\text{fusion}}_{h,w,t,c}, \tag{4}$$

where $0 \leq \gamma \leq 1$ is a residual parameter for the EEG features, and $\boldsymbol{\Phi} \in \mathbb{R}^{H\times W\times 1\times 1}$ is an FGA map that represents the spatial weight matrix extracted from fNIRS signals, obtained as follows:

$$\boldsymbol{\Phi} = \sigma_{\text{sig}}\left(\text{TAP}(f^{3\times 3\times 3@1}(\boldsymbol{F}^{\text{fnirs}}))\right), \tag{5}$$

where $f^{3\times 3\times 3@1}$ represents the convolutional layer with a filter size of $3 \times 3 \times 3$ and one output filter, which extracts the attention feature from the 3D fNIRS feature over the channel dimension. Thereafter, the temporal attentive pooling provides a large weight to temporally important time segments. The FGA map was not fixed and adaptively driven by every fNIRS

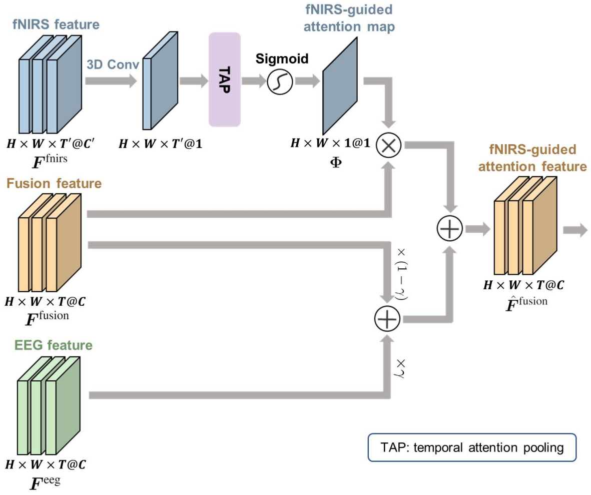


Fig. 6. The architecture of the fNIRS-guided attention layer.

input. The FGA map leads to reliable attention to spatially important regions, but unexpected low-weighted spatial attention may completely lose valuable information in the region. To mitigate the information loss of EEG features, we added EEG features of EEG branch to the FGA layer. The residual parameter γ is determined by the trainable parameter γ_{train} with sigmoid function σ_{sig}, that is, $\gamma = \sigma_{\text{sig}}(\gamma_{\text{train}})$. The overall structure of the FGA layer is shown in Fig. 6.

The fusion branch extracts joint representations from 3D EEG and fNIRS features. However, fNIRS signals have an inherent delay compared to EEG signals; thus, they can deteriorate the decoding performance at the beginning of trials. Therefore, to alleviate the performance degradation of the fusion branch, we utilized the prediction scores of both the EEG branch and the fusion branch. The final prediction was conducted by the weighted sum of the prediction scores of the EEG branch and fusion branch as follows:

$$\boldsymbol{y}^{\text{pred}} = \xi^{\text{eeg}}\boldsymbol{y}^{\text{eeg}} + \xi^{\text{fusion}}\boldsymbol{y}^{\text{fusion}}, \tag{6}$$

where $\boldsymbol{y}^{\text{eeg}}, \boldsymbol{y}^{\text{fusion}} \in \mathbb{R}^2$ are the prediction scores of the EEG and fusion branches. $\boldsymbol{\xi} = [\xi^{\text{eeg}}, \xi^{\text{fusion}}]$ is the prediction weight between two branches, which are obtained by

$$\boldsymbol{\xi} = \sigma_{\text{Soft}}\left(\text{FC}^{2\times1}\left([\boldsymbol{F}^{\text{eeg}}, \boldsymbol{F}^{\text{fusion}}]\right)\right), \tag{7}$$

where $\boldsymbol{F} \in \mathbb{R}^{64\times1}$ is the output of the last FC layer, and $[\cdot]$ is a concatenation function for the first dimension. The network parameters, such as kernel size, stride, and channel, of the fusion branch are equivalent to ESNet, described in Table I.

D. Loss Function for Training FGANet

Finally, we proposed a loss function for our proposed FGANet, which is divided into three parts: classification loss function L_{class}, fNIRS branch regularization L_{fnirs}, and FGA map regularization L_{fga}, as follows:

$$L = L_{\text{class}} + L_{\text{fnirs}} + \lambda L_{\text{fga}}, \tag{8}$$

where $\lambda > 0$ is a regularization parameter.

1) Classification Loss: The objective of this study was to decode brain activity from EEG and fNIRS signals. Therefore, we applied the cross-entropy function to the final prediction of FGANet as follows:

$$L_{\text{class}} = -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i \cdot \log(\boldsymbol{y}^{\text{pred}}(\boldsymbol{X}_i)) \tag{9}$$

where $\boldsymbol{y}_i$ is the label of the i-th input $\boldsymbol{X}_i = \{\boldsymbol{X}_i^{\text{eeg}}, \boldsymbol{X}_i^{\text{fnirs}}\}$, $\boldsymbol{y}^{\text{pred}}(\boldsymbol{X}_i)$ is the prediction score of FGANet for the input $\boldsymbol{X}_i$, '$\cdot$' denotes the dot product, and N is the number of input data.

2) fNIRS Branch Regularization: The fNIRS branch has the same structure as FSNet. Because the FGA layer essentially extracts the FGA map from fNIRS features of the fNIRS branch to identify spatially important regions, the performance significantly depends on how well fNIRS features represent exclusive patterns between classes. However, the classification loss function L_{class} is not sufficient to train the feature extractors of the fNIRS branch because the fNIRS feature is not directly used in the final classification. Hence, to accelerate the training process of the fNIRS branch, we add a cross-entropy loss function to minimize the classification accuracy of fNIRS data as follows:

$$L_{\text{fnirs}} = -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i \cdot \log(\boldsymbol{y}^{\text{fnirs}}(\boldsymbol{X}_i^{\text{fnirs}})). \tag{10}$$

where $\boldsymbol{y}^{\text{fnirs}}(\boldsymbol{X}_i^{\text{fnirs}})$ is the prediction score of the fNIRS branch for the input $\boldsymbol{X}_i^{\text{fnirs}}$.

3) FGA Map Regularization: The purpose of the FGA map $\boldsymbol{\Phi}$ is to highlight the spatially critical regions of the EEG feature based on the fNIRS feature to improve the classification performance. Despite EEG and fNIRS signals being highly correlated with spatial manners, a metric is required to measure the spatial correlation between EEG features and fNIRS features. Therefore, the Pearson correlation coefficient (PCC) is exploited to train the FGA map to maximize the correlation between two signals, defined as:

$$\text{PCC}(\boldsymbol{U}, \boldsymbol{V}) = \frac{\sum_{i,j}\left(U_{i,j} - \bar{U}\right)\left(V_{i,j} - \bar{V}\right)}{\sqrt{\sum_{i,j}\left(U_{i,j} - \bar{U}\right)}\sqrt{\sum_{i,j}\left(V_{i,j} - \bar{V}\right)}}, \tag{11}$$

where $\bar{U}, \bar{V}$ are the means of all the elements of the matrix $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{H\times W}$, respectively. Therefore, to regularize the FGA map, we maximize the PCC between the EEG feature and the FGA map as follows:

$$L_{\text{fga}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{3}\text{PCC}\left(\tilde{\boldsymbol{F}}^{\text{egg}(l)}(\boldsymbol{X}_i^{\text{eeg}}), \boldsymbol{\Phi}^{(l)}(\boldsymbol{X}_i^{\text{fnirs}})\right), \tag{12}$$

where

$$\tilde{F}_{h,w} = \frac{1}{TC}\sum_{t'=1}^{T}\sum_{c'=1}^{C}F_{h,w,t',c'}, \tag{13}$$

$\boldsymbol{F}^{\text{egg}(l)}, \boldsymbol{\Phi}^{(l)}$ represent the l-th layer of the 3D EEG feature and FGA map, respectively. To compress the 3D EEG feature with temporal and channel dimensions, the 3D EEG features were averaged to those dimensions.

However, FGA map regularization based on PCC can disturb the training of the EEG branch because the performance of fNIRS branch is significantly lower than that of EEG branch, and the fNIRS signals have an inherent delay to the stimuli. Therefore, we blocked the gradient flow to EEG branch oriented from the FGA map regularization to preserve the performance of the EEG branch.

IV. Result

A. Experimental Setup

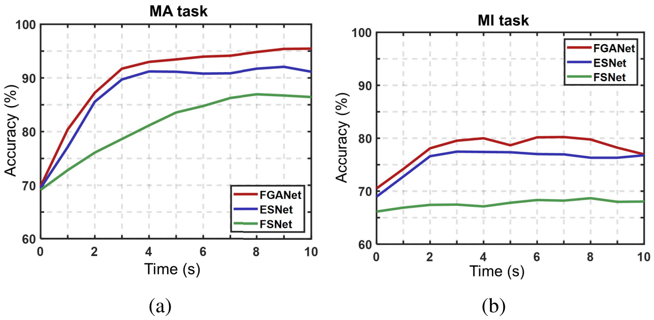
We evaluated our FGANet using a benchmark dataset recorded during MI ad MA tasks, as described in Section II. EEG and fNIRS data were divided into 3 s segments using a 1 s sliding window during the −2–10 s interval (including instruction and task periods) in Fig. 2, resulting in ten segments for each trial: −2–1 s, −1–2 s, . . . , 7–10 s windows. Let $(t - 3)$ s - t s time window be defined as the t s time segment. We conducted a 5-fold cross-validation for each subject and considered the average to obtain reliable results. The final result was obtained by averaging the results of all the subjects. To compare with various conventional algorithms, we calculated the mean and maximum accuracy for each algorithm. The mean accuracy was obtained by averaging the accuracies of ten segments in a trial, and the maximum accuracy was obtained by taking the maximum accuracy among ten segments in a trial.

The network parameters of the networks are summarized in Tables I and II. We utilized the Adam optimizer [53] to update the network parameters with a learning rate of 0.001 during 200 epochs. The initial trainable residual parameter γ_{train} was set to zero. We used the parameter fine-tuning process for optimizing the network parameters and regularization parameter λ. We first train the network using 90% data of the training set and then remaining 10% of the data is used as a validation set. The stride and kernel size of the first convolutional layer for the temporal dimension is optimized from the set (kernal, stride) $\in [(7, 2), (7, 4), (9, 2), (9, 4)]$ and regularization parameter is optimized from the set [1, 0.1, 0.01, 0.001]. Finally, the network parameter is determined as described in Tables I and II and regularization parameter λ is set to 0.1.

To calculate the mean accuracies of the comparison algorithms, we implemented the LDA algorithm [20] and the pth-PF algorithm [33], whose source codes were shared by the authors. We also implemented the pth-PF algorithm with EEG branch and fNIRS branch (pth-PF (EEG + fNIRS branch)) for a fair comparison with our method to state-of-the art algorithm because the feature extractor of pth-PF was built with a 1D CNN. Furthermore, we implemented the pth-PF algorithm with the EEG branch and fusion branch (pth-PF (EEG + fusion branch)) to compare the state-of-the art deep learning-based prediction algorithm with our weighted prediction method. The FC layers of ESNet and FSNet were utilized to fuse EEG and fNIRS signals in the 3D CNN + pth-PF algorithm.

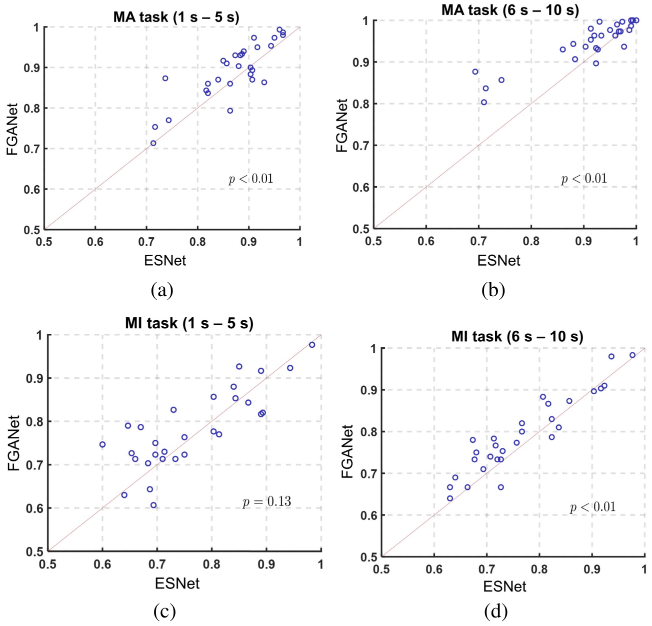
B. Performance Analysis on Change of Time Segments

Fig. 7. shows the classification result across the moving time window, where the x-axis indicates the right edge of the moving time window, and the y-axis shows the accuracy. Until the x-axis reached 3 s, part of the instruction period was included in the time segment (Fig. 2). Therefore, the performance of all the algorithms increased from 0 s to 3 s. After 3 s, the performance of ESNet was saturated, but that of FSNet continued to increase until 8 s, especially in the MA task. This is because the hemodynamic response is significantly slower than the electrical response, although it is not prominent in the MI task. Moreover, FGANet outperformed unimodal BCI systems (ESNet and FSNet) overall time segments for both MA and MI tasks.



(a) (b)

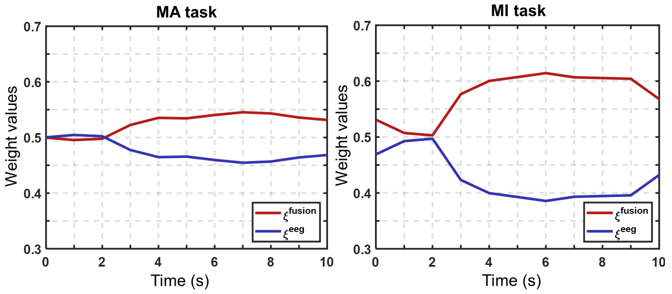
Fig. 7. Mean classification accuracy for the (a) MA task and (b) MI task at each 3 s time window in all trials (x-axis represents the right edge of the time window).



(a) (b) (c) (d)

Fig. 8. The mean classification accuracy for different time segments and task: MA task with averaging (a) 1 s to 5 s, and (b) 6 s to 10 s time segments, MI task with averaging (c) 1 s to 5 s, and (d) 6 s to 10 s time segments.

To show a statistically significant improvement in FGANet compared to ESNet, we calculated a paired t-test for the mean classification accuracy of ESNet and FGANet. Considering the delay in the hemodynamic response to the stimuli, we divided the time segments into the first half (1–5 s) and second half (6–10 s). As shown in Fig. 8(a) and (b), the performance of FGANet is significantly better ($p < 0.01$) than that of



(a) (b)

Fig. 9. The prediction weight values ξ^{fusion}, and ξ^{eeg} for the (a) MA task and (b) MI task with the 1 s moving window (x-axis represents the right edge of the moving window).

TABLE III

The Classification Accuracy (Mean ± Std) According to the Residual Parameter γ for MA and MI Task

Task	γ	Mean Acc (%)	Max Acc (%)
MA task	0	91.32 ± 06.59	94.60 ± 06.62
	0.3	91.24 ± 05.58	94.14 ± 06.79
	0.5	**92.29** ± 06.27	**95.69** ± 05.97
	0.7	91.66 ± 05.97	94.71 ± 05.47
	1.0	91.21 ± 05.99	94.37 ± 05.87
	γ_{train}	91.96 ± 05.82	95.46 ± 05.12
MI task	0	77.93 ± 08.66	79.60 ± 10.58
	0.3	78.36 ± 09.04	**80.29** ± 09.62
	0.5	**78.72** ± 08.50	**80.29** ± 10.28
	0.7	77.43 ± 08.86	78.74 ± 10.25
	1.0	77.94 ± 08.63	79.43 ± 09.61
	γ_{train}	78.59 ± 08.86	80.23 ± 09.63

ESNet in both groups of the MA task. In the MI task, the FGANet has a significantly better performance ($p < 0.01$) than ESNet in the second half group (6–10 s) (Fig. 8(d)), but there is no significant difference in performance between FGANet and ESNet at the beginning of the MI task (Fig. 8(c)). These results show that the proposed FGANet is a promising hybrid EEG-fNIRS BCI system that can outperform the EEG-standalone BCI system, but it has a limitation in improving the performance of the EEG-standalone BCI system when the performance of the fNIRS-standalone BCI system is significantly low to identify spatially important areas.

Fig. 9. shows the change in the prediction weight vector $\boldsymbol{\xi}$ according to the moving time window for the MA and MI tasks. In the figure, ξ^{fusion}, red line, is approximately 0.5 (for EEG and fusion predictions) at the beginning of the trial, and sharply increases after 2 s in both MA and MI tasks. The prediction weight of the fusion network ξ^{fusion} seems to reflect the delay in the hemodynamic response of fNIRS signals. These results show that the prediction weight can be adaptively adjusted according to the importance of the fNIRS signals.

C. Analysis of the Residual Parameter

In the FGA layer, the EEG feature was added to the fusion network to prevent the loss of EEG information. Table III summarizes the classification results for the change in the residual parameter γ for the EEG features. The results show that performance can vary according to the value of the residual parameter; the mean accuracy increases by more than 1.0%, and the maximum accuracy is enhanced by more than 1.5% at $\gamma = 0.5$ in both MI and MA tasks compared to the lowest performance. To analyze these results, we have rewritten the last FGA layer according to the residual parameter γ for the special case $\gamma = 0$ and 1 as follows:

$$\hat{F}^{\text{fusion}(3)}_{h,w,t,c} = F^{\text{fusion}(3)}_{h,w,t,c} + \Phi_{h,w,1,1} F^{\text{fusion}(3)}_{h,w,t,c}, \quad \text{if } \gamma = 0, \qquad (14)$$

and

$$\hat{F}^{\text{fusion}(3)}_{h,w,t,c} = F^{\text{eeg}(3)}_{h,w,t,c} + \Phi_{h,w,1,1} F^{\text{fusion}(3)}_{h,w,t,c}, \quad \text{if } \gamma = 1. \qquad (15)$$

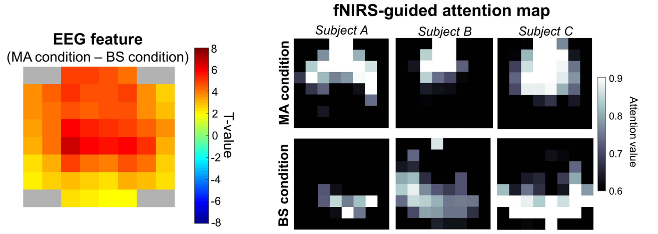
As shown in Eqs. (14) and (15), the EEG feature is not used in the fusion branch when $\gamma = 0$, whereas the previous fusion feature is not added to the FGA layer when $\gamma = 1$, where the performance is lower than when $\gamma = 0.5$. Therefore, a proper residual EEG feature ($0 < \gamma < 1$) can significantly increase performance by balancing the EEG information in the fusion branch. In particular, the performance of the FGA layer at only the last fusion feature ($\gamma = 1$) is lower than that of the stacked FGA layer ($0 < \gamma < 1$), which implies that early feature fusion has a higher potential to improve performance than the late feature fusion method.

The mean of trained residual parameters for all subjects is 0.54 and 0.61 for the MA and MI tasks, respectively (Table III). The mean and maximum accuracies of FGANet with the trained γ are 91.96% and 95.46% in the MA task, and 78.59% and 80.23% in the MI task, respectively. The performance of the trained residual parameter is comparable to the best accuracy of $\gamma = 0.5$, which obviates the need to tune the residual parameter for performance optimization.

D. Feature Visualization

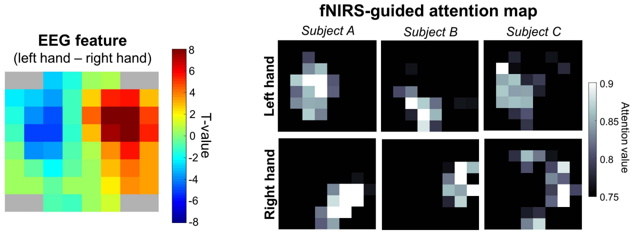
Figs. 10 and 11 show the feature visualization of the EEG feature $\tilde{\boldsymbol{F}}^{\text{fusion}(1)}$ as t-values and the FGA map $\boldsymbol{\Phi}^{(1)}$) for MA and MI tasks. Because the raw EEG feature is difficult to interpret, we used the t-value to find the discriminative region between the two classes. The t-value was calculated by paired t-test using all subjects. The red area of the EEG feature represents the region where the activation value of class 1 (mental arithmetic for MA task, left hand for MI task) is higher than that of class 2, and vice versa for the blue area. In the FGA map, a higher value (white color) represents an important region for brain decoding extracted from fNIRS signals.

In the MA task, the EEG feature of the MA condition was significantly higher than the baseline condition in all of the feature regions (red color in the MA task in Fig. 10(a)). This corresponds to the results of mental workload studies, which reported that the theta (4–8 Hz) power increases as the workload increases [20], [45]. In the EEG features, the t-value of the top area was higher than that of the bottom area for the MA task. This tendency is also reflected in the FGA map. As shown in Fig. 10(b), the FGA map also highlights the top area, whereas the bottom of the feature is in the baseline condition. This result shows that the FGA map properly highlights the discriminable region of the EEG feature.



(a) (b)

Fig. 10. The visualization of the (a) t-values (MA condition vs. BS condition) of EEG feature $\tilde{\boldsymbol{F}}^{\text{fusion}(1)}$, and the (b) fNIRS-guided attention map $\boldsymbol{\Phi}^{(1)}$ for the mental arithmetic task.



(a) (b)

Fig. 11. The visualization of the (a) t-values (left hand vs. right hand) of EEG feature $\tilde{\boldsymbol{F}}^{\text{fusion}(1)}$, and the (b) fNIRS-guided attention map $\boldsymbol{\Phi}^{(1)}$ for the motor imagery task.

In the MI task, the t-value of the EEG feature on the right side was positive, but the left area was negative (Fig. 11(a)). This result implies that the EEG feature value of the left area is higher when the subject imagines moving the left hand, while the right area is higher for the right hand. This result is similar to that of other studies, which reported that the motor imagery of the left hand was positively correlated with the left central lobe, while the right hand was positively correlated with the right central lobe caused by the inhibition process [20], [47]. Therefore, these results show that ESNet effectively compresses raw EEG signals into discriminable EEG features. The FGA map follows the trend of the t-values of the EEG features. The left side of the attention map was activated in the left-hand condition, whereas the right side was activated in the right-hand condition (Fig. 11(b)). This result shows that our FGA map can extract an important region for EEG decoding in the MI condition.

E. Ablation Study

Table IV presents an ablation study for the MA and MI tasks. The TAP layer can improve the mean accuracy in the MA task by almost 1%. In general, brain functioning for working memory tasks consists of two processes (manipulation and retention) and the retention period shows the significantly increased spectral power [52]. The experimental results show that the TAP layer can capture an underlying key feature of temporal neural dynamics in the retention period.

In addition, we investigated the effect of FGA map regularization on performance. The performance of FGANet without FGA map regularization L_{fga} is lower than that of FGANet with FGA map regularization in both MA and MI tasks. This indicates that the fusion performance can be improved by training the model parameters to reinforce the spatial correlation between the EEG and fNIRS signals. Therefore, it is crucial to extract the optimal joint representation for the underlying spatial correlation between the two signals to maximize the performance of the hybrid EEG-fNIRS BCI system.

TABLE IV

Classification Accuracy (Mean ± Std) According to the Ablation Study for MA Task and MI Task. Symbol "-" Denotes the Following Component Is Removed

Task	Model	Mean Acc (%)	Max Acc (%)
MA task	FGANet	**91.96** ± 05.82	**95.46** ± 05.12
	-TAP layer	91.02 ± 06.14	94.71 ± 06.09
	- residual parameter	91.32 ± 08.66	94.60 ± 10.58
	- FGA map regularization	90.81 ± 05.80	93.79 ± 05.27
	- weighted prediction	91.42 ± 06.07	94.71 ± 05.31
MI task	FGANet	**78.59** ± 05.82	**80.23** ± 05.12
	-TAP layer	78.12 ± 08.89	79.94 ± 10.88
	- residual parameter	77.93 ± 08.66	79.60 ± 10.58
	- FGA map regularization	76.58 ± 08.64	78.16 ± 10.03
	- weighted prediction	77.44 ± 08.95	79.77 ± 09.22

We demonstrated the superiority of the weighted prediction method as compared to other methods. The component "-weighted prediction" in Table IV represents the algorithm that extracts the prediction score by averaging the prediction scores of the EEG and fusion branches instead of using the weighted prediction method. The performance of "-weighted prediction" is inferior to that of FGANet. This is because the weighted prediction method can alleviate the performance degradation that is caused by the delayed hemodynamic response. As shown in Fig. 9, the prediction weight of the fusion branch is adaptively changed according to the reliability of the fusion branch. Therefore, the weighted prediction strategy is superior to the averaged prediction strategy.

F. Performance Comparison

We compared the variants of our proposed method with those of conventional algorithms, and the results are summarized in Table V. In conventional algorithms, the fNIRS-based BCI system proposed by Aydin *et al.* [57] achieved the best performance among unimodal BCI systems, and IDPF [25] is a state-of-the-art hybrid EEG-fNIRS BCI system that outperforms all other conventional algorithms. However, these methods used the entire 10 s data recorded during one trial for classification, whereas the deep learning-based fusion algorithm (pth-PF) [33] used only 3 s of data and predicted the class every second for real-time applications. Considering the difference in input size, it is difficult to compare the two algorithms fairly, but the performance of pth-PF can be regarded as the achievable maximum performance by deep learning approaches using 3 s data.

It is important to note that our ESNet algorithm using only EEG signals outperforms pth-PF in the MA task and the MI task. It seems that the MA task can be easily decoded by the 3D CNN structure compared to the MI task. This is because when we applied the EEG and fusion branch

TABLE V
THE CLASSIFICATION ACCURACY (MEAN ± STD) ACCORDING TO THE DIFFERENT ALGORITHM IN MA TASK AND MI TASK

Algorithm	Signal type	MA task		MI task	
		Mean Acc (%)	Max Acc (%)	Mean Acc (%)	Max Acc (%)
Ergun *et al.* [54]	EEG	-	88.71	-	-
Ergun *et al.* [55]	fNIRS	-	84.94	-	-
Ergun *et al.* [56]	fNIRS	-	-	-	72.36
Aydin *et al.* [57]	fNIRS	-	89.54	-	78.27
IDPF [25]	EEG+fNIRS	-	91.15	-	78.56
Shin *et al.*† [20]	EEG+fNIRS	75.60* ± 06.69	84.29* ± 09.07	60.91* ± 09.07	63.85* ± 10.83
*p*th-PF† [33]	EEG+fNIRS	87.24* ± 06.14	91.67* ± 06.09	75.90* ± 08.89	77.36* ± 10.88
*p*th-PF† (EEG+fNIRS branch)	EEG+fNIRS	87.95* ± 05.80	92.53* ± 05.27	73.10* ± 08.64	74.20* ± 10.03
*p*th-PF† (EEG+fusion branch)	EEG+fNIRS	87.99* ± 06.67	93.05* ± 05.89	74.67* ± 08.41	76.26* ± 09.88
ESNet	EEG	89.14* ± 07.73	92.07* ± 08.42	76.50* ± 09.63	77.47* ± 10.71
FSNet	fNIRS	82.34* ± 08.11	86.95* ± 09.19	67.80* ± 08.23	68.68* ± 09.00
FGANet	EEG+fNIRS	**91.96** ± 05.82	**95.46** ± 05.12	**78.59** ± 08.86	**80.23** ± 09.63

1) "†" represents the performance of our implementation model for the conventional algorithms.
2) "*" represents the significantly differences ($p < 0.05$, paired t-test) compared to the FGANet. The paired t-test is only conducted on the our implementation model.

used in the proposed algorithm to the conventional *p*th-PF algorithm, the maximum accuracy increased by 0.86% in the MA task, but decreased by 3.16% in the MI task. Therefore, a high-dimensional input structure is not always effective in improving the performance.

However, the proposed fusion method, FGANet, outperformed the state-of-the-art algorithm (IDPF) in both the MA and MI tasks. More specifically, the maximum accuracy of FGANet is 4.3% and 1.7% higher than that of the IDPF in the MA and MI tasks, respectively. Furthermore, the mean accuracy of FGANet was greater than 2% as ($p < 0.05$) compared to the *p*th-PF, and *p*th-PF (EEG + fNIRS branch), which is the state-of-the-art deep learning-based algorithm in both the MA and MI tasks. These results show that our fNIRS-guided fusion method can considerably improve the performance of the hybrid EEG-fNIRS BCI system, and is applicable to real-time applications. Furthermore, compared to *p*th-PF (EEG + fusion branch), the mean and max accuracy of FGANet was significantly higher ($p < 0.05$). This result shows the superiority of our prediction method as compared to the conventional algorithms.

V. CONCLUSION

In this study, we proposed the fNIRS-guided attention network (FGANet) as a deep learning-based early fusion structure. First, the 1D multi-channel EEG and fNIRS signals were converted into 3D EEG and fNIRS tensors to spatially align the EEG and fNIRS signals. Thereafter, we extracted a joint representation of both signals using the proposed FGA layer. In the FGA layer, fNIRS features were used to create the FGA map that identifies the important regions of the EEG features for reliable EEG decoding. The FGA map was trained to maximize the spatial correlation between the EEG features and the FGA map using FGA map regularization. Finally, the prediction score of the EEG branch was added to the final prediction to alleviate the performance deterioration caused by the inherent delay of fNIRS signals. The experimental results showed that our FGANet outperformed ESNet, FSNet, and the state-of-the-art fNIRS-EEG fusion method. Furthermore, we verified that the FGA map properly highlighted the spatially important regions of the EEG features. This framework can be extensively applied to any neural network for hybrid fNIRS-EEG BCIs.

REFERENCES

[1] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1034–1043, Jun. 2004.

[2] G. Purtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001.

[3] J. Wolpaw *et al.*, "Brain-computer interface technology: A review of the first international meeting," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 164–173, Feb. 2000.

[4] J.-H. Jeong, K.-H. Shim, D.-J. Kim, and S.-W. Lee, "Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1226–1238, May 2020.

[5] Z. T. Al-Qaysi, B. B. Zaidan, A. A. Zaidan, and M. S. Suzani, "A review of disability EEG based wheelchair control system: Coherent taxonomy, open challenges and recommendations," *Comput. Methods Programs Biomed.*, vol. 164, pp. 221–237, Oct. 2018.

[6] R. Scherer, G. R. Müller, C. Neuper, B. Graimann, and G. Pfurtscheller, "An asynchronously controlled EEG-based virtual keyboard: Improvement of the spelling rate," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 979–984, Jun. 2004.

[7] S. Ramgopal *et al.*, "Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy," *Epilepsy Behav.*, vol. 37, pp. 291–307, Aug. 2014.

[8] A. T. Tzallas *et al.*, "Automated epileptic seizure detection methods: A review study," in *Epilepsy-Histological, Electroencephalographic and Psychological Aspects*, Feb. 2012, pp. 75–98.

[9] C. Melissant, A. Ypma, E. E. E. Frietman, and C. J. Stam, "A method for detection of Alzheimer's disease using ICA-enhanced EEG measurements," *Artif. Intell. Med.*, vol. 33, no. 3, pp. 209–222, Mar. 2005.

[10] S. Fazli *et al.*, "Enhanced performance by a hybrid NIRS–EEG brain computer interface," *NeuroImage*, vol. 59, no. 1, pp. 519–529, 2012.

[11] J. Mellinger *et al.*, "An MEG-based brain–computer interface (BCI)," *NeuroImage*, vol. 36, no. 3, pp. 581–593, Jul. 2007.

[12] S. H. Sardouie and M. B. Shamsollahi, "Selection of efficient features for discrimination of hand movements from MEG using a BCI competition IV data set," *Frontiers Neurosci.*, vol. 6, p. 42, Apr. 2012.

[13] M. A. Tanveer, M. J. Khan, M. J. Qureshi, N. Naseer, and K.-S. Hong, "Enhanced drowsiness detection using deep learning: An fNIRS study," *IEEE Access*, vol. 7, pp. 137920–137929, 2019.

[14] T. K. K. Ho, J. Gwak, C. M. Park, and J. Song, "Discrimination of mental workload levels from multi-channel fNIRS using deep leaning-based approaches," *IEEE Access*, vol. 7, pp. 24392–24403, 2019.

[15] J. Kwon and C.-H. Im, "Performance improvement of near-infrared spectroscopy-based brain-computer interfaces using transcranial near-infrared photobiomodulation with the same device," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2608–2614, Dec. 2020.

[16] L. G. Lim *et al.*, "A unified analytical framework with multiple fNIRS features for mental workload assessment in the prefrontal cortex," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2367–2376, Nov. 2020.

[17] S.-S. Yoo *et al.*, "Brain–computer interface using fMRI: Spatial navigation by thoughts," *NeuroReport*, vol. 15, no. 10, pp. 1591–1595, Jul. 2004.

[18] G. Rota, G. Handjaras, R. Sitaram, N. Birbaumer, and G. Dogil, "Reorganization of functional and effective connectivity during real-time fMRI-BCI modulation of prosody processing," *Brain Lang.*, vol. 117, no. 3, pp. 123–132, Jun. 2011.

[19] G. Pfurtscheller, "The hybrid BCI," *Frontiers Neurosci.*, vol. 4, p. 3, Apr. 2010.

[20] J. Shin *et al.*, "Open access dataset for EEG+NIRS single-trial classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1735–1745, Oct. 2017.

[21] J. Shin, J. Kwon, and C.-H. Im, "A ternary hybrid EEG-NIRS brain-computer interface for the classification of brain activation patterns during mental arithmetic, motor imagery, and idle state," *Frontiers Neuroinform.*, vol. 12, p. 5, Feb. 2018.

[22] F. Al-Shargie, T. B. Tang, and M. Kiguchi, "Stress assessment based on decision fusion of EEG and fNIRS signals," *IEEE Access*, vol. 5, pp. 19889–19896, 2017.

[23] L.-W. Ko *et al.*, "Multimodal fuzzy fusion for enhancing the motor-imagery-based brain computer interface," *IEEE Comput. Intell. Mag.*, vol. 14, no. 1, pp. 96–106, Feb. 2019.

[24] C.-H. Han, K.-R. M´uller, and H.-J. Hwang, "Enhanced performance of a brain switch by simultaneous use of EEG and NIRS data for asynchronous brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 10, pp. 2102–2112, Oct. 2020.

[25] X. Jiang, X. Gu, K. Xu, H. Ren, and W. Chen, "Independent decision path fusion for bimodal asynchronous brain–computer interface to discriminate multiclass mental states," *IEEE Access*, vol. 7, pp. 165303–165317, 2019.

[26] S. Fazli, J. Mehnert, J. Steinbrink, and B. Blankertz, "Using NIRS as a predictor for EEG-based BCI performance," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 4911–4914.

[27] H. Morioka *et al.*, "Decoding spatial attention by using cortical currents estimated from electroencephalography with near-infrared spectroscopy prior information," *NeuroImage*, vol. 90, pp. 128–139, Apr. 2014.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[30] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[31] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[32] A. M. Chiarelli, P. Croce, A. Merla, and F. Zappasodi, "Deep learning for hybrid EEG-fNIRS brain–computer interface: Application to motor imagery classification," *J. Neural Eng.*, vol. 15, no. 3, Apr. 2018, Art. no. 036028.

[33] Z. Sun, Z. Huang, F. Duan, and Y. Liu, "A novel multimodal approach for hybrid brain–computer interface," *IEEE Access*, vol. 8, pp. 89909–89918, 2020.

[34] Y. Kwak, K. Kong, W.-J. Song, B.-K. Min, and S.-E. Kim, "Multilevel feature fusion with 3D convolutional neural network for EEG-based workload estimation," *IEEE Access*, vol. 8, pp. 16009–16021, 2020.

[35] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3D convolutional neural network for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2164–2177, Oct. 2019.

[36] J.-H. Jeong, K.-H. Shim, D.-J. Kim, and S.-W. Lee, "Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1226–1238, May 2020.

[37] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, Nov. 2005, pp. 399–402.

[38] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Proc. Workshop Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, Sep. 2014, pp. 474–490.

[39] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Multilevel sensor fusion with deep learning," *IEEE Sensors Lett.*, vol. 3, no. 1, pp. 1–4, Jan. 2019.

[40] Y.-R. Cho, S. Shin, S.-H. Yim, K. Kong, H.-W. Cho, and W.-J. Song, "Multistage fusion with dissimilarity regularization for SAR/IR target recognition," *IEEE Access*, vol. 7, pp. 728–740, 2019.

[41] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *Proc. ACM Multimedia Conf.*, Oct. 2016, pp. 978–987.

[42] C. S. Roy and C. S. Sherrington, "On the regulation of the blood-supply of the brain," *J. Physiol.*, vol. 11, nos. 1–2, pp. 85–158, Jan. 1890.

[43] P. Lachert, D. Janusek, P. Pulawski, A. Liebert, D. Milej, and K. J. Blinowska, "Coupling of Oxy- and deoxyhemoglobin concentrations with EEG rhythms during motor task," *Sci. Rep.*, vol. 7, no. 1, pp. 1–9, Nov. 2017.

[44] M. Takeuchi *et al.*, "Brain cortical mapping by simultaneous recording of functional near infrared spectroscopy and electroencephalograms from the whole brain during right median nerve stimulation," *Brain Topography*, vol. 22, no. 3, pp. 197–214, Aug. 2009.

[45] A. Gundel and G. F. Wilson, "Topographical changes in the ongoing EEG related to the difficulty of mental tasks," *Brain Topography*, vol. 5, no. 1, pp. 17–25, 1992.

[46] M. Benedek, R. J. Schickel, E. Jauk, A. Fink, and A. C. Neubauer, "Alpha power increases in right parietal cortex reflects focused internal attention," *Neuropsychologia*, vol. 56, pp. 393–400, Apr. 2014.

[47] L. Brinkman, A. Stolk, H. C. Dijkerman, F. P. de Lange, and I. Toni, "Distinct roles for Alpha- and beta-band oscillations during mental simulation of goal-directed actions," *J. Neurosci.*, vol. 34, no. 44, pp. 14783–14792, Oct. 2014.

[48] L. Kocsis, P. Herman, and A. Eke, "The modified beer–lambert law revisited," *Phys. Med. Biol.*, vol. 51, no. 5, pp. N91–N98, Mar. 2006.

[49] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.

[50] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2012.

[51] 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, Jan. 2014.

[52] O. Jensen, J. Gelfand, J. Kounios, and J. E. Lisman, "Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task," *Cerebral Cortex*, vol. 12, no. 8, pp. 877–882, 2002.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[54] E. Ergün and O. Aydemir, "A new evolutionary preprocessing approach for classification of mental arithmetic based EEG signals," *Cognit. Neurodyn.*, vol. 14, no. 5, pp. 609–617, Apr. 2020.

[55] E. Ergün and O. Aydemir, "Decoding of binary mental arithmetic based near-infrared spectroscopy signals," in *Proc. 3rd Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2018, pp. 201–204.

[56] E. Ergun and O. Aydemir, "Classification of motor imaginary based near-infrared spectroscopy signals," in *Proc. 26th Signal Process. Commun. Appl. Conf. (SIU)*, May 2018, pp. 1–4.

[57] E. A. Aydin, "Subject-specific feature selection for near infrared spectroscopy based brain-computer interfaces," *Comput. Methods Programs Biomed.*, vol. 195, Oct. 2020, Art. no. 105535.