A Context-Supported Deep Learning Framework for Multimodal Brain Imaging Classification

Jianmin Jiang, Ahmed Fares, and Sheng-Hua Zhong

***Abstract*—Over the past decade, "content-based" multimedia systems have realized success. By comparison, brain imaging and classification systems demand more efforts for improvement with respect to accuracy, generalization, and interpretation. The relationship between electroencephalogram (EEG) signals and corresponding multimedia content needs to be further explored. In this paper, we integrate implicit and explicit learning modalities into a context-supported deep learning framework. We propose an improved solution for the task of brain imaging classification via EEG signals. In our proposed framework, we introduce a consistency test by exploiting the context of brain images and establishing a mapping between visual-level features and cognitive-level features inferred based on EEG signals. In this way, a multimodal approach can be developed to deliver an improved solution for brain imaging and its classification based on explicit learning modalities and research from the image processing community. In addition, a number of fusion techniques are investigated in this work to optimize individual classification results. Extensive experiments have been carried out, and their results demonstrate the effectiveness of our proposed framework. In comparison with the existing state-of-the-art approaches, our proposed framework achieves superior performance in terms of not only the standard visual object classification criteria, but also the exploitation of transfer learning. For the convenience of research dissemination, we make the source code publicly available for downloading at GitHub (https://github.com/aneeg/dual-modal-learning).**

***Index Terms*—Deep learning, electroencephalogram (EEG), explicit learning modality, implicit learning modality, object classification.**

Manuscript received September 15, 2018; revised January 8, 2019; accepted February 24, 2019. Date of publication April 2, 2019; date of current version November 21, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61620106008, in part by the the Natural Science Foundation of Guangdong Province under Grant 2016A030310053, in part by the Shenzhen Emerging Industries of the Strategic Basic Research Project under Grant JCYJ20160226191842793, in part by the Shenzhen high-level overseas talents program, in part by the National Engineering Laboratory for Big Data System Computing Technology, and in part by the Inlife-Handnet Open Fund. This paper was recommended by Associate Editor J. Han. *(Jianmin Jiang and Ahmed Fares are co-first authors.) (Corresponding author: Sheng-Hua Zhong.)*

J. Jiang and S.-H. Zhong are with the Research Institute for Future Media Computing, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China (e-mail: jianmin.jiang@szu.edu.cn; csshzhong@szu.edu.cn).

A. Fares is with the Research Institute for Future Media Computing, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Department of Electrical Engineering, the Computer Engineering branch, Faculty of Engineering at Shoubra, Benha University, Cairo 2900, Egypt (e-mail: ahmed.fares@szu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2904615

I. Introduction

HUMAN brain analytics has long been researched across a number of communities, including neural science, brain science, psychology, etc. At present, the brain interface is primarily examined via two approaches, functional magnetic resonance imaging (fMRI) and electroencephalograms (EEGs), where an EEG is a recording of voltage fluctuations produced by ionic current flows in the neurons of the brain [1]. While reflecting the brain's spontaneous electrical activities, an EEG also has the potential to provide a subjective response based on individual experiences [2], [3]. As a noninvasive brain signaling technique, an EEG provides high spatiotemporal resolution data, presenting a vivid reflection of the dynamics of the brain [4], which makes it ideal for a variety of research fields, such as brain-computer interface (BCI) [5]–[7], affective state recognition [8], [9], and diagnosis of brain-related diseases, such as epilepsy [10], [11], Alzheimer's [12], [13], Parkinson's etc. [14].

For the past decades, an understanding of EEG data evoked by specific stimuli/objects has been the primary goal of BCI and other important EEG-related research fields. Hence, EEG-based object/event classification becomes a key component across all these communities. Further, a number of psychological and neuroscience studies have demonstrated that up to a dozen special object categories can be classified by the event-related potential (ERP) recorded through EEGs [15]–[17], such as human faces. With applications of machine learning, a range of models have been developed [18]–[20] to address the problem of classifying visual objects via EEGs. However, most stimuli/objects in these studies are designed with only a single object set inside a clean background because enormous ambiguity surrounds the interpretation of EEG data, and multichannel EEG data sequences are generally only available in small quantities. In addition, EEGs are high dimensional yet have low signal-to-noise ratios, and the differences among individual subjects incur considerable temporal and spatial variability [21]. One study that is similar to ours in terms of the categorization objective is reported in [22], in which Walther *et al.* proposed an approach to estimate the categories of natural scenes using fMRI for only six categories. Specifically, they used fMRI and distributed patterns to analyze what regions of the brain can classify natural scenes. In practice, fMRI showed great potential in the brain imaging classification process; however, its main disadvantage lies in the experimental costs. This limitation is overcome by lower-cost techniques, such as EEGs, which provide higher temporal-resolution data compared to fMRI, but are susceptible to the aforementioned problems represented by lower signal-to-noise

ratios and spatial resolutions, posing significant challenges for brain imaging classification. With the progress of machine learning, some important limitations of the traditional neural networks have been overcome [23]–[27]. Inspired by such advancements, content-based multimedia understanding has achieved remarkable success in object detection [28] and image scene classification [29]. In contrast to the success of content-based multimedia understanding over the recent years, research on EEGs is still limited, providing an enormous scope for considerable improvement in terms of information extraction and classification of EEGs, including accuracy, generalization, and interpretability.

At present, image classification based on physiological signal analysis (implicit analysis) and content-based multimedia analysis (explicit analysis) have been two independently active research areas, and the relationship between the massive amount of physiological signals and the corresponding multimedia content has been relatively unexplored [30]. Nevertheless, the information from these two sources is likely to be complementary. On the one hand, physiological signal data may help us better understand the process of image classification inside human brains, and thus, it can be helpful for designing a highly robust brain-inspired model to classify the visual objects under complex backgrounds and occlusions. On the other hand, the feature extraction methods proposed for multimedia content analysis could inspire us to discover new and unstudied physiological signal patterns for developing more powerful and more intelligent object classification algorithms. In other words, brain-based image classification could be significantly improved if we can simultaneously leverage both the implicit and explicit modalities in designing the classification algorithms. Our extensive literature survey indicates, however, that there is no existing work integrating them for EEG-based image classification, although multimedia content analysis has achieved impressive progress over the past decade.

In this paper, we propose a novel deep brain analytics framework together with a multimodal approach for EEG-based image classification by integrating implicit and explicit modalities. Specifically, a consistency test based on a mapping between image content features and EEG-based features is added to promote potential solutions, and better performances are achieved compared to the state-of-the-art methods for EEG-based object classifications. Further, our proposed deep framework also demonstrates a good generalization capability in object categorization over a number of publicly available and widely adopted benchmarking datasets. In comparison with the existing research efforts and the corresponding state-of-the-art methods, the novelty of our contributions can be highlighted as follows. 1) We introduce a new concept of integrated implicit learning and explicit learning modalities to provide an alternative solution for the problem of brain imaging classification. 2) We propose a new deep brain analytics framework to exploit not only the strength of integrated multiple modalities, but also the advantages of the added consistency test for recommending potential targets and the fusion of individual classification results. 3) We carry out extensive experiments, and the results demonstrate that our proposed deep framework achieves superior performances in comparison with the existing state-of-the-art approaches.

II. Related Work

In psychology, stimuli refer to objects or events that cause a sensory or behavioral response in an organism. Therefore, stimuli form the basis of perception and behavior for human brain analytics, which has been intensively researched across the areas of neural science, psychology, and neural computation. Researchers aspire to present, analyze, distinguish, and understand how the human brain receives, handles, and processes rich and varied information in the real world through EEG signals, among which information about visual content and emotions is the primary target for research and analysis. Therefore, multimedia data containing a large amount of visual content information and emotional information are considered to be extremely suitable stimuli material, which are widely used in the acquisition, analysis, and classification of EEG signals [31], [32]. The research on multimedia content computing and multimedia emotional computing based on EEG signals has attracted enormous attention across relevant research communities [19], [32]–[35].

Before the popularity of deep learning methods, the primary approaches for image classification were predominantly feature based, and the commonly used features mainly included time-frequency features extracted by signal analysis methods, such as the power spectral density [36], bandpower [37], independent component analysis (ICA) [38], and differential entropy [39].

With the extensive application and in-depth promotion of deep learning, an ever-increasing number of brain and neuroscience research teams are exploiting its strength in designing ambitious algorithms to achieve intelligent understanding and perceptional analysis of brain activities via EEGs or fMRI. In [16], deep belief networks and deep automatic encoders to resolve the ERP P300 and non-P300 signals were reported. In [40], Yin and Zhang proposed a single-channel EEG classification method with a deep belief network to evaluate mental workload and mental fatigue states. In [20], an SVM classifier was trained to classify visually evoked EEG data according to 12 different object categories. In [41], a frequential deep belief network (FDBN) for classification tasks in motor imagery and adaptive EEG analysis was proposed. In [42], Gogna *et al.* proposed deep learning methods to solve the problem of reconstruction and classification of EEG data. In [43], a four-layer convolutional neural network to detect interictal discharges from intracranial EEG data was described, and determination of the effects of convolutional neural networks on decoding and visualization of EEGs was attempted and reported [44]. In [45], Dong *et al.* used the rectified linear unit activation function and long short-term memory (LSTM) on time frequency domain features to classify sleep stages. In [46], Stober *et al.* used CNNs and an autoencoder to classify audio-evoked EEG recordings. In [47], a compact full convolutional network (EEGNet) was proposed and applied to four different brain-machine interface classification tasks. In [48], Spampinato *et al.* used long-term and short-term memory network learning to obtain an EEG data representation based on image stimuli and constructed a mapping relationship from natural image features

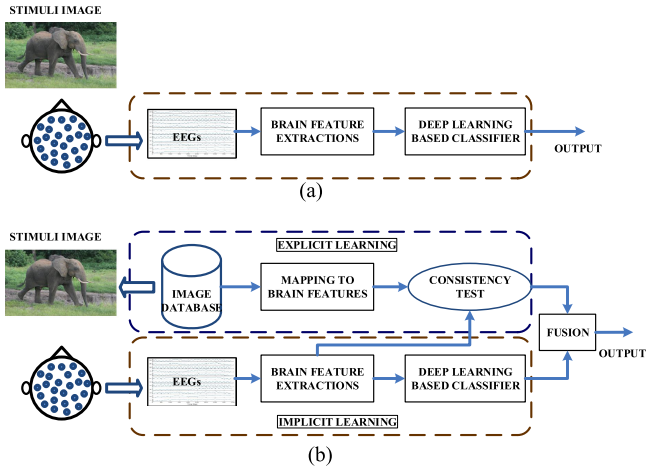

Fig. 1. (a) Illustration of the existing research on the single modality of implicit learning. (b) Illustration of our proposed framework with dual modalities, implicit learning, and explicit learning.

to EEG characterization. Finally, they used the new representation of EEG signals for classification of natural images. Compared with traditional methods, these deep learning based approaches have achieved outstanding results in realizing their respective research objectives. While these methods have demonstrated the capability of using brain signals and deep learning for classification purposes, none of them simultaneously integrated implicit and explicit modalities, and the state-of-the-art classification accuracy achieved to date by Spampinato *et al.* was 82.9% [48], leaving significant space for further research and improvement.

III. Proposed Multimodal Classification Algorithm

As shown in Fig. 1(a), the existing efforts on brain image classification are primarily limited to a single modality, so-called implicit learning, where EEG or fMRI signals are directly processed to extract brain features for learning and classification. Whether the learning process is designed as conventional or deep learning based, the essential strategy of a single modality remains unchanged. While such single modality strategies have achieved good progress across areas of both image processing and computer vision, the rich source of image content from which stimuli are selected for producing EEG or fMRI signals is basically ignored. To exploit the great successes, especially those achieved by deep learning based approaches toward intelligent image content analysis and classification, we introduce as a new strategy a second modality, so-called explicit learning, as shown in Fig. 1(b), which is added to target the rich source of images used for stimuli and to determine whether their analysis could provide further assistance in improving the classifications of EEGs.

Given m classes of images $\mathbf{G} = \{\mathbf{g}_i\}_{i=1}^{m}$, from which both training images and testing images are selected as stimuli to produce EEG signals, we apply deep learning based networks to extract brain feature vectors $\mathbf{B}_i \in \mathbb{R}^{d_b}$, where d_b stands for the dimension of the brain features. To create the second modality and exploit the rich source of the image database for improved brain signal classification, we propose to examine the m classes of images inside the original image database and determine if any of their content description and analysis in the so-called explicit learning can boost the implicit learning-based classification. To minimize the computing cost and the algorithm complexity, we select a number of representative images out of each class to characterize all the images inside the corresponding class, i.e., $\mathbf{g}_i, i \in [1, m]$. As the widely researched approach of clustering proves to be powerful in characterizing images, we apply a pixel-based clustering method to cluster all the images within each individual class such that the centroid of each cluster is taken as the most representative image for its corresponding class. In this way, the K-means clustering takes each class of images $\mathbf{g}_i$, as input and produces the most representative images per class, $\mathbf{G^r} = \{\mathbf{g}_i^{\mathbf{r}}\}_{i=1}^{m}$, as output, where $\left\{\mathbf{r}_1^i, \mathbf{r}_2^i, \ldots, \mathbf{r}_{n_i}^i\right\} \in \mathbf{g}_i^{\mathbf{r}}$ declares that each representative image class $\mathbf{g}_i^{\mathbf{r}}$ actually contains n_i representative images. In our algorithm design, we simply use the number of centroids as the number of representative images since the clustering is applied to all images of each individual class. While $\mathbf{G}$ denotes the original image set, with $\mathbf{g}_i$ being the ith class of images, $\mathbf{G^r}$ denotes the extracted representative image database, with $\mathbf{g}_i^{\mathbf{r}}$ being the ith class of representative images.

As seen in Fig. 1, the implicit learning modality essentially relies on deep learning of brain features to determine which class the input brain feature is associated with. To this end, the brain features of all the training images can be regarded as providing a certain level of ground truth for describing all the different classes. To allow a certain level of flexibility and tolerance for those images that could be selected as the test image, we cluster the entire database rather than the training images exclusively to produce the representative class for the consistency test and hence propose to add a second modality, i.e., the so-called explicit learning, by directly mapping all the representative images into the brain feature space and hence construct an explicit brain feature for each of them. In other words, we do not actually extract the brain feature from the EEG signals of those representative images; rather, we derive the brain feature directly from the mapping process because not all of the features have EEGs available. We then carry out a similarity-based consistency test to determine which representative brain feature provides the closest match for the brain feature of the stimuli image, and thus, the corresponding class can be selected as the recommended classification output to assist the classification from the implicit learning modality, the details of which are described as follows.

Let $\mathbf{B^t} = \left\{\mathbf{b}_1^{\mathbf{t}}, \mathbf{b}_2^{\mathbf{t}}, \ldots, \mathbf{b}_{d_b}^{\mathbf{t}}\right\} \in \mathbb{R}^{d_b}$ be the brain feature vector of the test image and $\mathbf{B}_{ij}^{\mathbf{r}} = \left\{\mathbf{B}_{i1}^{\mathbf{r}}, \mathbf{B}_{i2}^{\mathbf{r}}, \ldots, \mathbf{B}_{in_i}^{\mathbf{r}}\right\}$ be the brain feature set for the ith class of representative images; we calculate the distance between $\mathbf{B^t}$ and $\mathbf{B}_{ij}^{\mathbf{R}}$ as follows:

$$d\left(\mathbf{B^t}, \mathbf{B}_{ij}^{\mathbf{R}}\right) = \frac{1}{d_b}\sum_{k=1}^{d_b}\left(\mathbf{b}_k^{\mathbf{t}} - \mathbf{b}_{ij}^{k}\right)^2 \tag{1}$$

where $\mathbf{b}_{ij}^{k}$ is the kth element of the jth brain feature vector $\mathbf{B}_{ij}^{\mathbf{r}}$ inside the ith class of representative images.

The index of the representative image, for which the mapped brain feature of the input test image is the closest, can be derived via the following:

$$(i', j') = \underset{i\in[1,m], j\in[1,n_i]}{\operatorname{argmin}} d\left(\mathbf{B^t}, \mathbf{B}^{\mathbf{r}}_{ij}\right). \tag{2}$$

To ensure that such recommended class candidates have the best possible opportunity to include the true classification result, we allow a certain level of tolerance by applying (2) not only to the first minimum value, but also to the second and third minimum values. As the ith class contains n_i representative images, it is likely that a number of images across different classes are within the inclusion of the minimum match. If this is the case, all the brain features are integrated and averaged as a new brain feature that is then sent to the implicit learning model for reclassification, as if the brain features of the selected representative images were extracted from the EEG signals. Given that concatenation fusion is widely used in the machine learning community, examples of which include the inception model of GoogLeNet [49], applications in multimodal deep learning [50], etc., we propose to concatenate the input brain feature $\mathbf{B^t}$ with the brain features $\mathbf{B^r}$ from representative images to complete the reclassification. Such classified results are referred to as $\mathbf{C}_2$, and the direct classification of EEGs via implicit learning is referred to as $\mathbf{C}_1$. On the other hand, if all the index values derived by (2) are within a single class, this class candidate will be directly used as the recommended classification result, which is referred to as $\mathbf{C}_3$. Details of such a recommendation test are summarized as follows:

$$\gamma_c = \begin{cases} \mathbf{C}_3 & \text{if } (i', j') \in \mathbf{B}^{\mathbf{r}}_{i'} \forall_{i',j'} \\ \mathbf{C}_2 = \varphi\left(\dfrac{\alpha}{\eta}\displaystyle\sum_{k=1}^{\eta}\left\{\mathbf{B}^{\mathbf{r}}_{i'j'}\right\}_k + \beta\mathbf{B^t}\right) & \text{else} \end{cases} \tag{3}$$

where γ_c stands for the recommended class, η is the total number of brain features that achieve the minimum distance for $\mathbf{B^t}$ via (2), and α and β are the two weighting coefficients balancing the contributions of $\mathbf{B^r}$ and $\mathbf{B^t}$, which are determined via empirical study and training. Finally, $\varphi(\cdot)$ stands for the deep learning based classification obtained via the implicit learning modality.

To make the final decision for the classification output and integrate all individual classification results, $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3\}$, we add a simple fusion as follows:

$$\mathbf{C} = \begin{cases} \mathbf{C}_1 & \text{if } \mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}_3 \\ C_3 & \text{else if } P(\mathbf{C}_2) < T \bigcap P(\mathbf{C}_1) < T \\ \bar{C} & = \underset{\mathbf{C}\in[\mathbf{C}_1,\mathbf{C}_2]}{\operatorname{argmax}} \left(\delta P(\mathbf{C}_1), P(\mathbf{C}_2)\right) \text{ else} \end{cases} \tag{4}$$

where T is a threshold, which is determined empirically during the training process, $P(\mathbf{C})$ is the probability that $\mathbf{C}$ is the correct classification, and δ is an adjustment coefficient designed to constrain or boost $P(\mathbf{C}_1)$.

Essentially, (4) is designed to integrate all the classification results produced by the multimodal information fusion (two modalities), in which the first choice is straightforward and the second choice states that if neither $\mathbf{C}_1$ nor $\mathbf{C}_2$ has sufficient probability to justify its output, we would directly adopt the recommended classification as the output. Finally, the third choice states that the output is the one corresponding to the maximum probability among $P(\mathbf{C}_1)$ and $P(\mathbf{C}_2)$.

IV. Deep Learning Based Multimodal Framework

A. System Overview

Fig. 2 shows an overview of our proposed deep framework. As seen, the purpose of the real image is two-fold: As the input for the explicit learning and as the stimuli for evoking brain signals. The integrated EEG-based brain image classification consists of four stages, i.e., feature encoding, EEG regression, consistency test, and information fusion.

In the modality of implicit learning, the information is first extracted from raw EEG signals to construct brain cognitive features via an LSTM network, and these features are then fed into a number of later stages inside the framework, including the EEG regression stage, the consistency test stage, and the information fusion stage.

In the modality of explicit learning, on the other hand, the most representative images of each class are obtained via the clustering technique, and information is extracted from the image to construct visual features via the CNN, which is referred to as the feature encoding stage. To introduce the explicit learning into our framework, we propose to apply a KNN regression process and map the encoded visual feature into an EEG description, paving the way for the consistency test and hence producing recommendations for potential classification candidates $(\mathbf{C}_3)$ as described in (4).

The consistency test plays a gap-bridging role between the two modalities of implicit learning and explicit learning, where the brain features from EEGs and the content features from images are comparatively tested to estimate their consistency and fulfil the integration of the two modalities. Following that, a fusion stage is added to optimize the collective considerations of these individual classification results and hence deliver the best possible final classification performances. Under this circumstance, our fusion design is critical to ensure that both false positives and false negatives can be significantly reduced.

B. Feature Encoding

The feature encoding stage aims at extracting the brain cognitive feature representation $\mathbf{B}$ and the visual feature $\mathbf{V}$ from raw EEG signals and the input images, respectively.

In preparing the modality of implicit learning, the raw EEG signals are processed through a recurrent module, in which an LSTM is used as an encoder, and the temporal sequence is projected into a feature space $\mathbf{B} \in \mathbb{R}^{d_b \times n}$. Specifically, the LSTM-based encoder network consists of a single LSTM layer and the regular nonlinear output layer. It takes EEG brain signals as input and produces the brain cognitive features representation $\mathbf{B}$ as output.

To prepare the modality of explicit learning for context support and complementary classification, we send all representative images to a CNN-based GoogLeNet [49] encoder to

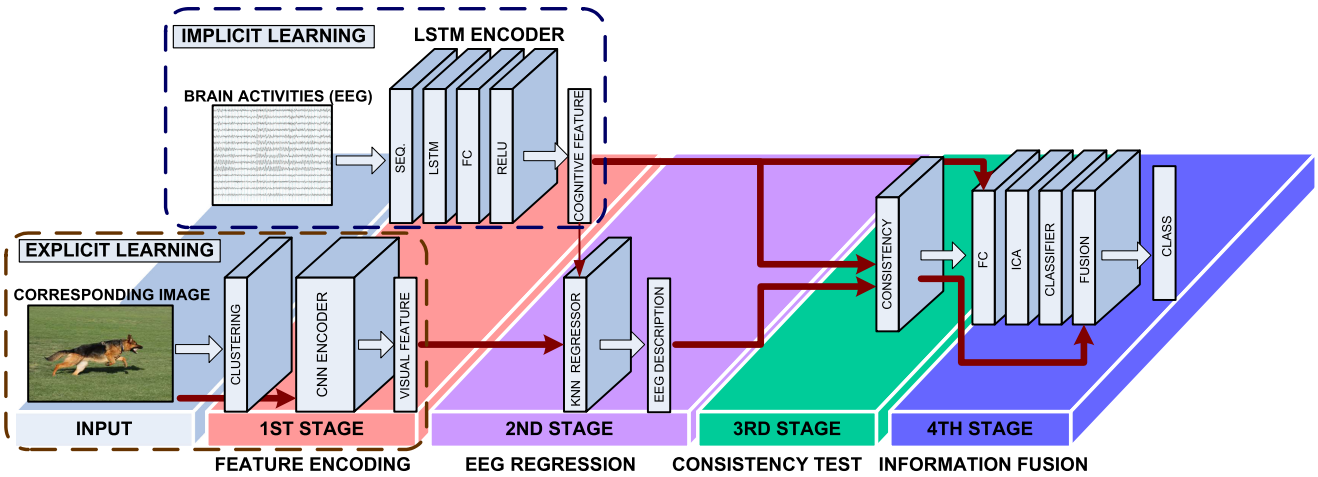

Fig. 2. Structural illustration of our proposed deep framework.

produce a set of visual feature vectors aimed toward mapping the visual content into EEG-compatible brain features. The overall structure and the relationship among all individual elements are shown in Fig. 2.

C. EEG Regression

To complete the mapping from the visual content representation to the brain feature description and ensure that the two different modalities, implicit learning and explicit learning, are compatible for integration, we add an EEG regression stage to project the visual features of the most representative images per class onto the brain cognitive features $\mathbf{B}$ via the visual features $\mathbf{V}$, producing the EEG description of the most representative images $\mathbf{B}_{ij}^{\mathbf{r}}$.

The KNN regressor layer has three inputs and one output. The three inputs include the visual feature of the images $\mathbf{V}$, the visual feature of the representative images $\mathbf{V}^{\mathbf{r}}$, and the brain cognitive feature $\mathbf{B}$. The output includes the regressed EEG description of the most representative images $\mathbf{B}_{ij}^{\mathbf{r}}$. The KNN regressor layer compares the visual features of the representative images to those of the images, and the K-nearest images to the representative images are retrieved. After that, the mean of the brain cognitive features associated with the K-returned images is calculated and considered to be the EEG description or characterization of the most representative images $\mathbf{B}_{ij}^{\mathbf{r}}$.

D. Consistency Test

As it is known that EEGs are often degraded by noise interference, leading to the possibility that their classifications could be less reliable, we propose a consistency test to overcome this problem and help reduce the nonreliability. Given that stimuli images presented to human subjects for extracting EEG sequences are bounded in the 40 classes inside ImageNet-EEG [48], we extract n_i most representative images from each class via clustering techniques, and use these representative images to perform a consistency test on the brain activity analysis result, i.e., the classified output from the implicit learning modality, and to determine which representative images it is consistent with. As a result, the corresponding class can be taken as an alternative classification result, and we expect that the alternative classification should be the same as that from the implicit learning and classification. For those results that differ from each other, we apply a further fusion stage to finalize the classification output.

Essentially, the aim of the third stage is to produce the EEG representation of the consistent set based on the representative images. On the one hand, the first input of the consistency test is the brain cognitive feature vector, which is derived from EEG monitoring of the brain activities $\mathbf{B}^{\mathbf{t}}$. On the other hand, the second input is the EEG description of the most representative images $\mathbf{B}_{ij}^{\mathbf{r}}$, which is derived from the EEG regression stage. In this way, it is guaranteed that both $\mathbf{B}^{\mathbf{t}}$ and $\mathbf{B}_{ij}^{\mathbf{r}}$ are consistency testable. In addition, the KNN similarity measure is utilized to check the similarity between the inputs ($\mathbf{B}^{\mathbf{t}}$ and $\mathbf{B}_{ij}^{\mathbf{r}}$) and produce the K-nearest EEG representations of each image in the test set based on the EEG description of the representative images. After that, the mean of the EEG representation associated with the K representative images is calculated and considered as the EEG description of the consistent set. Consequently, either the classified results are verified by the consistency test or alternative classifications are produced for further fusion and integration.

E. Information Fusion

The information fusion stage consists of three main parts, including ICA, the classifier layer, and the fusion layer. This stage plays a constructive role in improving the classification accuracy for the proposed deep framework.

ICA is placed before the layer of classifiers as a feature selection module, which takes the EEG features B from the LSTM network as input and returns the independent statistical features as output. We implement the reconstruction ICA objective function based on the work reported in [51]. After ICA, two classifiers have been investigated, including the SoftMax classifier and the multiclass support vector machine (SVM).

To optimize the individual classification results derived by the consistency test and by direct classification of EEG signals, we add a multimodal information fusion layer, attempting to exploit both their individual strengths and their complementary advantages. Regarding the specific multimodal information fusion algorithm design, our fusion problem contends with two classification results corresponding to two modalities. On the one hand, we have direct classification results from the first stage, and on the other, we have the consistency test results, which could contradict each other in principle. Therefore, a question arises: What classification result should we adopt as the final one in order to maximize the classification accuracy?

Two main approaches have been investigated in our work, including decision-based multimodal fusion and feature-based multimodal fusion. In the decision-based multimodal fusion, we test the probability-based approach, and in the feature-based multimodal fusion, we further test three techniques, concatenation fusion, SUM, and MAX fusion.

For the probability-based fusion, we use the output probability of the classifier layer from both modalities, where the first modality output is $\mathbf{C}_1$ and the second modality outputs are $\mathbf{C}_2$ and $\mathbf{C}_3$, to determine the final classification result according to (4). Following the strategy given in (4), the second modality outputs work as a rectifier to improve the final classification accuracy.

V. Experiments

To evaluate our proposed deep framework and the introduced concept of context-supported multimodal learning, we conduct three phases of experiments. In the first phase, we try to evaluate the EEG-based object classification performance for our proposed deep framework on the publicly available dataset ImageNet-EEG [48]. In the second phase, we measure the generalization capability of the proposed framework on a subset of the visual classification dataset Caltech-101 [52]. By generalization capability, we mean how our proposed deep framework performs if it is applied to classify those objects or images that have not been seen before. In the third phase, the proposed deep framework is tested under a transfer learning setup. To benchmark our approach with multimodalities, we first compare the proposed framework with the existing state-of-the-art methods to verify its effectiveness. Then, we conduct additional evaluations to explore the performance of the proposed framework in more detail.

A. Experimental Settings and Training Details

Our experiments are conducted on two datasets, including the EEG-based classification dataset ImageNet-EEG and a subset of the visual-based classification dataset Caltech-101. ImageNet-EEG is a publicly available EEG dataset for brain imaging classification proposed by Spampinato *et al.* [48]. Caltech-101 includes 17 classes that coincidentally have the same names as those in ImageNet-EEG. Hence, those images are selected to construct the subset utilized to evaluate the generalization capability of our deep framework. For benchmarking purposes, the proposed deep framework is compared with the EEG-based object classification method [48], which is the latest research work published in 2017 on the same dataset. We also perform comparisons with several of the latest deep learning models for visual-based object classification, including AlexNet [23], VGGNet-16 [53], VGGNet-19 [53], GoogLeNet [49], and ResNet-101 [24].

TABLE I
Classification Performance Comparison Between Our Proposed Deep Framework and the State-of-the-Art Method [48]

Models	Accuracy
Proposed deep framework	94.1%
RNN-based model [48]	82.9%

In the modality of implicit learning, the iteration limit is set to 200 and the batch size is set to 440 for the parameters of the LSTM encoder in the first stage of the proposed deep framework. In the modality of explicit learning, K is set to 3 for the parameters of the pixel-based clustering and the feature-based clustering in the first stage of the proposed deep framework. In the third stage of the proposed deep framework, concerning the parameters for the consistency test, the number of nearest neighbours K is set to 3. In the fourth stage of the proposed deep framework, the number of extracted features from ICA is set to 70, and the iteration limit is set to 400. Our method is implemented on the Tesla P100 GPU.

As KNN and clustering employ an unsupervised learning mode and the CNN is pretrained on ImageNet for visual feature extraction, all three modules, the KNN regressor, the clustering, and the CNN, do not need to be trained. The LSTM encoder is trained on the EEG data with their labels. For the KNN regressor, which attempts to map the features into a user-specific EEG space, both features at the input are visual features, one for representative images and the other for training images. For the consistency test, both inputs for KNN are cognitive features from the representative images and training images.

B. EEG-Based Object Classification

In the first phase of experiments, we try to validate the effectiveness of our deep framework for EEG-based object classification. Our experiments are tested on the standard EEG dataset ImageNet-EEG. As ImageNet-EEG is collected using a 128-channel cap with active, low-impedance electrodes (actiCAP 128Ch), it includes the EEG signals of six subjects produced by asking them to look at visual stimuli, which are images selected from a subset of ImageNet [54], containing 40 classes with 50 images in each class. During the experiment, each image was shown on the computer screen for 500 ms.

Table I summarizes the experimental results in terms of the classification accuracies for both our proposed deep framework and the existing state-of-the-art method reported in [48]. As seen, while the precision rate achieved by our proposed deep framework is 94.1%, the existing state-of-the-art comparison is 82.9%.

To quantify the contribution of each stage designed in our proposed deep framework, we further carried out experiments to explore the effectiveness of different configurations of the

TABLE II
COMPARATIVE ASSESSMENT OF THE PROPOSED FRAMEWORK UNDER DIFFERENT CONFIGURATIONS

			Framework											
	Configurations		1	2	3	4	5	6	7	8	9	10	11	12
1ST stage:	Feature encoding	Pixel-based	✓		✓		✓		✓	✓	✓	✓	✓	✓
		Feature-based		✓		✓		✓						
4TH stage:	Classifier	SoftMax	✓	✓										
		SVM			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Feature selection	ICA					✓	✓		✓		✓		✓
	Decision-based multi-modal fusion	Probability-based	✓	✓	✓	✓	✓	✓						
	Feature-based multi-modal fusion	Concatenation							✓	✓				
		SUM									✓	✓		
		MAX											✓	✓
Accuracy	%		89.5	89.7	92.5	92.3	94.1	94.1	87.8	89.5	87	92.2	86.5	92.2

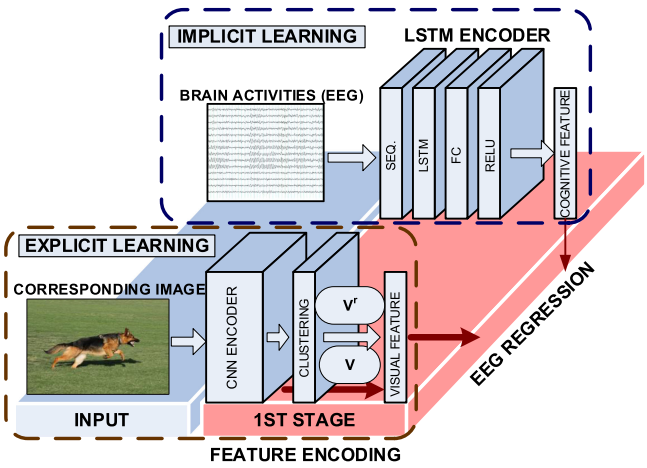

Fig. 3. Illustration of the alternative feature-based clustering.

individual stages. For the clustering stage, an alternative consideration is feature based, in which all the deep features of images inside each class can be clustered instead of their pixels, and then, the centroids are taken as the most representative deep features for their corresponding class (see Fig. 3). For the fusion stage, individual elements considered include 1) with or without ICA; 2) selection of different classifiers, including SoftMax and SVM; and 3) choice of different fusion methods, including probability based, corresponding to the pixel-based clustering, and feature based, corresponding to the feature-based clustering. Under the feature-based fusion, we could directly concatenate features as described in (3) or add every individual element of the features together as the fusion method (SUM). Finally, we could also select the most influential feature via $\mathrm{MAX}\{\mathbf{B^t}, \mathbf{B^r}\}$ (MAX).

Table II reports the experimental results in terms of the classification precision rates for all the configurations, from which we can observe and draw a number of conclusions, as described below.

First, the feature-based clustering in the first stage is better than the pixel-based clustering method, although the improvement is limited. This occurs if we select SoftMax as the classifier in the fourth stage of the proposed framework (see Fig. 2). These results are demonstrated in Table II by configurations 1 and 2.

Second, the performance of the pixel-based clustering is similar to (or even better than) the performance of the feature-based clustering if we select SVM as the classifier in the fourth stage. These results are demonstrated in Table II by configurations 3 to 6. These results illustrate why we select the pixel-based clustering method in the first stage when using SVM as the classifier in the fourth stage.

Third, we find that the SVM classifier is always better than the SoftMax classifier in the fourth stage. These results are demonstrated by configurations 1 to 4 in Table II. While the best performance of the SoftMax classifier is 89.7% (configuration 4), the best performance of the SVM classifier is 92.5% (configuration 5).

Fourth, we find that the performance of ICA plus SVM is always better than using SVM alone in the fourth stage. If we use ICA to reduce the feature dimension, the performance is always better, as demonstrated by configurations 3–6 in Table II. While the best performance of the SVM implementation is 92.5% (configuration 3), the best performance of employing ICA plus SVM is 94.1% (configurations 5 and 6).

Fifth, in the feature-based fusion method, SUM and MAX are better than concatenation fusion in the fourth stage. This is demonstrated by configurations 7–12 in Table II. While the best performance of the concatenation fusion method is 89.5% (configuration 8), the best performance of the SUM/MAX fusion method is 92.2% (configuration 10). The reason that the SUM/MAX fusion method outperforms the concatenation-based fusion is mainly due to the nature of the inputs to the fusion function. As the inputs are consistent features, the model-free fusion, SUM/MAX, is more suitable, while the concatenation fusion requires universal approximation to estimate the model parameters.

Sixth, the probability-based fusion method is better than the feature-based fusion. This is demonstrated by configurations 1–12 in Table II. While the best performance of the feature-based fusion method is 92.2% (configurations 10–12), the best performance of the probability-based fusion method is 94.1% (configurations 5 and 6).

Finally, we can conclude that the addition of the explicit learning modality does help the implicit learning modality achieve better performances for EEG-based object classification. Without the explicit learning modality, the best performance of implicit learning alone is only 90.5% when SVM is used as

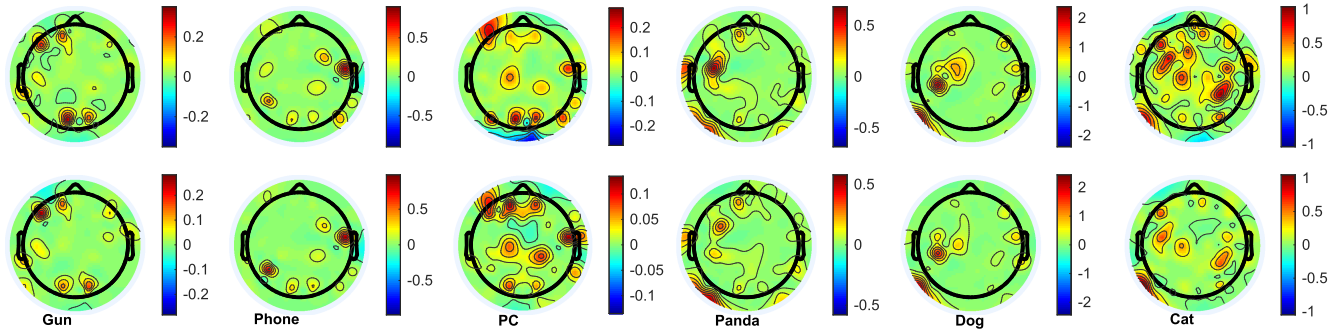

Fig. 4. Scalp distribution of the average energy for all participants and sessions for six categories, including "gun," "phone," "desktop PC," "panda," "dog," and "cat."

the classifier, and the best performance of the implicit learning alone is 82.9% when SoftMax is used as the classifier (due to space limitations, we do not list these results in Table II). With the addition of the explicit learning modality, however, the best performance of these settings is 94.1% (configuration 5) and 89.7% (configuration 2).

One novelty of our proposed framework is to integrate explicit and implicit learning modalities with a similarity-based consistency test on the representative images for every category. While the representative images are selected via pixel-level clustering, whether these representative images will truly trigger particular responses at the brain level for their corresponding classes remains questionable. To answer this question, we provide an average energy distribution of six categories for all participants and all sessions (first row) and the average energy distribution of the most representative images (second row) in Fig. 4. As seen, the neural activations of the representative images are close to that of all images in the same category, and an interesting observation is that the activations of three objects ("gun," "phone," and "desktop PC") are different from that of three animals ("panda," "dog," and "cat"). As three adorable animals, the activations of "panda," "dog," and "cat" share some level of similarity, as seen in Fig. 4, especially in the temporal area, indicating a strong sensitivity to visual perceptions, such as animal faces. In the ImageNet-EEG dataset, which includes 40 classes, most of the categories are not related to human emotions. From the scalp distribution of the average energy for all participants across all sessions, however, it is obvious that there exist higher responses at prefrontal areas, and these EEG data could be used for emotion classification.

C. Generalization Test for the Proposed Deep Framework

To test our proposed deep framework for its generalization capability in classifying brain images that are not previously seen by the framework via EEGs, we carry out the second phase of experiments on another widely used dataset, Caltech-101 [52]. Caltech-101 has 17 classes that are named the same as those in ImageNet-EEG, which creates an opportunity for us to carry out the generalization test by using the corresponding EEG signal sequences provided in [48]. For the convenience of result analysis and comparative studies, we construct a subset with all 17 classes from Caltech-101 to implement the second experiment, and the 17 classes include airplanes, bass, butterfly, camera, car with side view, cellphone, chair, cup, Dalmatian (dog), electric guitar, elephant, grand piano, lotus, panda, pizza, revolver, watch, and wildcat. The total number of images is 2059, and the number of images for each class is 121 on average.

TABLE III
COMPARATIVE GENERALIZATION TEST BETWEEN OUR PROPOSAL AND THE EXISTING STATE-OF-THE-ART METHOD [48]

Models	Accuracy
Proposed deep framework	80%
CNN-based model [48]	77%

Specifically, images from the 17 classes are taken as the input for GoogLeNet, and the output of the last fully connected layer is used as the extracted visual features. To maintain the necessary compatibility between the extracted visual features (explicit learning) and the brain cognitive features (implicit learning) for a smooth integration of the two modalities, we project all these visual features onto the brain feature space via the learned KNN regression module as shown in Fig. 2, and this is implemented without performing training on any image in Caltech-101.

The experimental results are summarized in Table III, which lists the classification performances for both the proposed deep framework and the existing benchmark [48]. As seen, our proposed framework outperforms the benchmark by 3%, indicating the following: 1) our proposed framework has a better generalization capability in classifying visual objects not previously seen, and 2) human visual capabilities can be learned and exchanged via machine learning.

For the convenience of further analysis and comparative investigation, Fig. 5 presents the confusion matrix of each category for Caltech-101. The rows represent the 17 classes from Caltech-101, and the first 17 columns represent the corresponding classes in ImageNet-EEG. The last column is used to represent the rest of the 40 classes from ImageNet-EEG. As we directly use the learned models trained via ImageNet-EEG to predict the images in Caltech-101, we find that some images are incorrectly classified into classes that do not exist in Caltech-101. In the confusion

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Others
1	73	7	0	0	10	0	0	0	0	0	0	0	1	0	0	0	0	9
2	3	85	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	6
3	3	0	31	0	5	0	0	0	0	0	0	0	1	0	1	1	8	50
4	6	0	2	91	0	0	0	0	0	0	0	0	0	0	0	0	0	2
5	8	3	0	0	66	0	0	0	0	0	0	0	0	0	0	0	0	24
6	0	0	0	0	3	84	0	0	0	0	0	0	0	0	0	0	0	14
7	0	0	0	0	0	0	56	0	0	19	0	0	3	0	0	0	0	22
8	2	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	0	4
9	0	0	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	5
10	0	0	0	0	0	0	0	0	0	81	0	0	1	0	0	0	0	17
11	3	0	0	0	5	0	0	0	0	0	65	0	0	0	4	0	0	23
12	2	0	0	0	0	0	2	0	0	2	0	61	21	0	0	0	0	13
13	1	0	0	0	0	0	0	0	0	0	0	0	94	0	0	0	0	5
14	0	0	0	0	0	0	8	0	0	4	0	0	0	58	0	0	0	30
15	0	0	0	0	1	0	0	0	0	1	0	0	0	0	94	0	0	4
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	0	8
17	2	0	2	0	5	0	0	0	0	0	0	0	0	0	0	0	80	12

Fig. 5. Confusion matrix, where the rows represent the 17 classes from Caltech-101, and the first 17 columns represent the corresponding classes in ImageNet-EEG. The column "Others" represents the rest of the 40 classes from ImageNet-EEG.

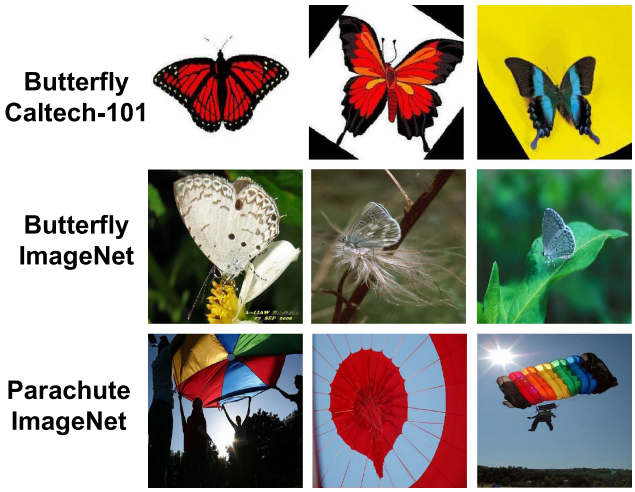

Fig. 6. Sample images from Caltech-101 and ImageNet.

matrix, these samples contribute to the confusion values labelled as "Others."

As seen in Fig. 5, the classification accuracy of the "butterfly" category is worse than others, only 31%. Additional examination even indicates that most of the images from this category are wrongly classified as "parachute" by our framework, which prompted our further investigation into the results. Additionally, Fig. 6 shows several sample images from the class "butterfly" in both Caltech-101 and ImageNet-EEG. As seen, the visual content from the "butterfly" images of Caltech-101 is obviously very different from that from the "butterfly" images of ImageNet-EEG. In other words, although the two classes are named the same, their images do not have any similarity in terms of visual objects. In contrast, the visual content from the "butterfly" images of Caltech-101 is very similar to that from the "parachute" images of ImageNet-EEG.

D. Transfer Learning via the Proposed Deep Framework

To improve the classification performances of our proposed deep framework on Caltech-101, we add a training process by selecting images from the 17 classes of Caltech-101. We do not want to change the EEG sequences inside ImageNet-EEG [48]; however, the power of transfer learning can be exploited to augment the EEG-based classification with help from the training process via images from Caltech-101.

As shown in Fig. 2, the essential integration of the two different modalities of explicit learning and implicit learning is supported because the compatibility between the visual features directly extracted from images and the brain cognitive features extracted from the EEGs is preserved. To this end, we establish an indirect mapping from the visual level to the brain cognitive level by transfer learning. In the first round of indirect mapping, we aim to obtain the brain cognitive level representation of each image in the training set of the 17 classes in Caltech-101. Specifically, the visual features extracted from the Caltech-101 images are further processed by KNN-based regression and then represent the brain cognitive features. By comparing the extracted visual features to those of the training images in ImageNet-EEG, the nearest neighboring images from ImageNet-EEG can be retrieved, and the mean of the brain cognitive features associated with these returned images are taken as the brain cognitive-level representation of each image in the training set of Caltech-101. In the second round of indirect mapping, our goal is to obtain the brain cognitive-level representation of the test images in the 17

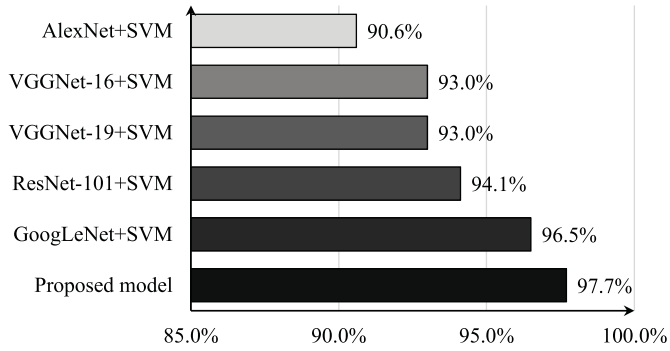

Fig. 7. Classification performance comparison between our proposed deep framework and the latest deep learning methods as the feature extractors, including AlexNet, VGGNet-16, VGGNet-19, GoogLeNet, and ResNet-101, where SVM is used as the classifier.

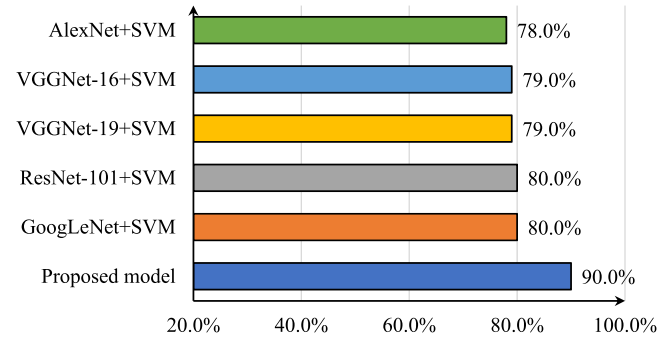

Fig. 8. Classification performance comparison between our proposed deep framework and the latest deep learning methods, including AlexNet, VGGNet-16, VGGNet-19, GoogLeNet, and ResNet-101, on the category of "butterfly".

classes of Caltech-101. The procedure is very similar to the first round, and the only difference is that the regression is carried out between the test set and the training set of Caltech-101 rather than between the training set of Caltech-101 and ImageNet-EEG. After the indirect mapping has been constructed, the brain cognitive-level representations of the training images in the 17-class Caltech-101 set are used to train the SVM classifier, and those of the test images are used to evaluate our proposed deep framework.

For the purpose of maintaining fair comparisons, AlexNet [23], VGGNet-16 [53], VGGNet-19 [53], GoogLeNet [49], and ResNet-101 [24], which are pretrained by ImageNet, are used as the feature extractors, and SVM is used as the classifier. The subset of the 17 classes in Caltech-101 is split into training, validation, and test sets, with percentages of 80%, 10%, and 10%, respectively. For all of those compared, images from the training set in the 17 classes of Caltech-101 are used to train SVM, the validation set is utilized to determine the parameters of SVM, and we evaluate the performance on the test set.

The experimental results are summarized in Fig. 7, from which we can see that the classification accuracy of our proposed framework is better than those of all the other approaches. Considering that our proposed deep framework primarily relies on classification of EEG sequences in ImageNet-EEG and the training with images from Caltech-101 is only exploited via the power of transfer learning, our proposed deep framework still remains competitive, particularly since the benchmark, GoogLeNet+SVM, directly exploits the training process with images from Caltech-101, rendering our proposal at a significant disadvantage.

For the convenience of further comparative analysis, we specifically focus on the category of "butterfly" and carry out a detailed investigation of the performance in the generalization test. As expected, the performance on "butterfly" is not good because the visual appearances of the images across the two datasets are significantly different and uncorrelated. As a result, a further analysis is conducted to study the performances after the training procedure. Fig. 8 demonstrates the classification accuracies on the category of "butterfly" achieved by our proposed deep framework and the latest deep learning methods, including AlexNet, VGGNet-16, VGGNet-19, GoogLeNet, and ResNet-101. As seen, the classification accuracy of our proposed deep framework is better than those of all the other methods. While the "butterfly" classification accuracy achieved by our deep framework is 90%, the "butterfly" classification accuracies achieved by AlexNet, VGGNet-16, VGGNet-19, GoogLeNet, and ResNet-101 are 78%, 79%, 79%, 80%, and 80%, respectively.

TABLE IV
PERFORMANCE EVALUATION OF THE PROPOSED DEEP FRAMEWORK FOR TWO DIFFERENT CONFIGURATIONS

	Framework	
Configurations	3	5
Accuracy	96.5 %	97.7 %

To assess the influence of different configurations, we further carry out experiments with a range of configurations to evaluate how the classification performances vary, and Table IV summarizes the top two results. As seen, configuration 5, which achieves the highest precision rate in the EEG-based object classification experiment, also achieves the highest precision rate and outperforms all the others.

VI. CONCLUSION

In this paper, by integrating implicit and explicit learning modalities, we propose a novel deep framework for EEG-based brain imaging classification. Our proposed framework provides an improved solution for the problem that, given an image used to stimulate brain activities, we should be able to identify which class the stimuli image comes from by analyzing the prompted EEG signals. As the visual cognitive capability of human brains is primarily researched via fMRI across the neural science and brain cognitive computing communities, significant challenges exist for processing EEG sequences, as they are exposed to noise and a high level of ambiguity exists. To address these challenges, we exploit the explicit learning widely researched across the areas of computer vision, digital media, and machine learning. To this end, we add a consistency test between the cognitive features extracted from EEG sequences and the visual features extracted from representative images to produce alternative and improved

solutions. Extensive experiments support that our proposed approach outperforms the existing state-of-the-art methods under various contexts and set-ups. The success achieved not only indicates that EEGs have significant potential for capturing brain activities for visual cognitive computing, but also opens up a new direction for explicit learning that, while widely researched in computer science, could also play significant roles in understanding and exploring the human brain.

A number of possibilities can be identified for further research, including applications of our deep framework for EEG-based brain activity understanding and interpretation, and testing the reliability and robustness of our framework toward recognition of visual objects and content inside human brains.

REFERENCES

[1] L. Yuan and J. Cao, "Patients' EEG data analysis via spectrogram image with a convolution neural network," in *Intelligent Decision Technologies*, I. Czarnowski, R. J. Howlett, and L. C. Jain, Eds. Cham, Switzerland: Springer, 2018, pp. 13–21.

[2] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image Vision Comput.*, vol. 31, no. 2, pp. 164–174, 2013.

[3] E. Kroupi, A. Yazdani, J.-M. Vesin, and T. Ebrahimi, "EEG correlates of pleasant and unpleasant odor perception," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 1s, 2014, Art. no. 13.

[4] C. Guger, A. Schlogl, C. Neuper, D. Walterspacher, T. Strein, and G. Pfurtscheller, "Rapid prototyping of an EEG-based brain-computer interface (BCI)," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 9, no. 1, pp. 49–58, Mar. 2001.

[5] A. M. Green and J. F. Kalaska, "Learning to move machines with the mind," *Trends Neurosciences*, vol. 34, no. 2, pp. 61–75, 2011.

[6] D. Wu, "Online and offline domain adaptation for reducing BCI calibration effort," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 4, pp. 550–563, Aug. 2017.

[7] Y. Mishchenko, M. Kaya, E. Ozbay, and H. Yanar, "Developing a 3- to 6-state EEG-based brain-computer interface for a virtual robotic manipulator control," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 4, pp. 977–987, Apr. 2019.

[8] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.

[9] B. Hu, X. Li, S. Sun, and M. Ratcliffe, "Attention recognition in EEG-based affective learning research using CFS+KNN algorithm," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 1, pp. 38–45, Jan. 2018.

[10] M. Fan and C. Chou, "Detecting abnormal pattern of epileptic seizures via temporal synchronization of EEG signals," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 3, pp. 601–608, Mar. 2019.

[11] D. Wang *et al.*, "Epileptic seizure detection in long-term EEG recordings by using wavelet-based directed transfer function," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 11, pp. 2591–2599, Nov. 2018.

[12] Z. Song, B. Deng, J. Wang, and R. Wang, "Biomarkers for Alzheimer's disease defined by a novel brain functional network measure," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 1, pp. 41–49, Jan. 2019.

[13] D. Labate, F. L. Foresta, G. Morabito, I. Palamara, and F. C. Morabito, "Entropic measures of EEG complexity in Alzheimer's disease through a multivariate multiscale approach," *IEEE Sensors J.*, vol. 13, no. 9, pp. 3284–3292, Sep. 2013.

[14] X. Chen, X. Chen, R. K. Ward, and Z. J. Wang, "A joint multimodal group analysis framework for modeling corticomuscular activity," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1049–1059, Aug. 2013.

[15] K. Das, B. Giesbrecht, and M. P. Eckstein, "Predicting variations of perceptual performance across individuals from neural activity using pattern classifiers," *Neuroimage*, vol. 51, no. 4, pp. 1425–1437, 2010.

[16] J. Kulasingham, V. Vibujithan, and A. De Silva, "Deep belief networks and stacked autoencoders for the p300 guilty knowledge test," in *Proc. IEEE EMBS Conf. Biomed. Eng. Sci.*, 2016, pp. 127–132.

[17] F. Li *et al.*, "Deep models for engagement assessment with scarce label information," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 4, pp. 598–605, Aug. 2017.

[18] J. Wang, E. Pohlmeyer, B. Hanna, Y.-G. Jiang, P. Sajda, and S.-F. Chang, "Brain state decoding for rapid image retrieval," in *Proc. 17th ACM Int. Conf. Multimedia.*, 2009, pp. 945–954.

[19] J. Moon, Y. Kwon, K. Kang, C. Bae, and W. C. Yoon, "Recognition of meaningful human actions for video annotation using EEG based user responses," in *Proc. Int. Conf. Multimedia Model.*, 2015, pp. 447–457.

[20] B. Kaneshiro, M. P. Guimaraes, H.-S. Kim, A. M. Norcia, and P. Suppes, "A representational similarity analysis of the dynamics of object processing using single-trial EEG classification," *PLoS ONE*, vol. 10, no. 8, p. e0135697, 2015.

[21] S. Stober, "Learning discriminative features from electroencephalography recordings by encoding similarity constraints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 6175–6179.

[22] D. B. Walther, C. Eamon, F. F. Li, and D. M. Beck, "Natural scene categories revealed in distributed patterns of activity in the human brain," *J. Neuroscience Official J. Soc. Neuroscience*, vol. 29, no. 34, pp. 10573–10581, 2009.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

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

[25] Z. Zhang, Y. Lu, L. Zheng, S. Li, Z. Yu, and Y. Li, "A new varying-parameter convergent-differential neural-network for solving time-varying convex QP problem constrained by linear-equality," *IEEE Trans. Autom. Control*, vol. 63, no. 12, pp. 4110–4125, Dec. 2018.

[26] Z. Zhang *et al.*, "A varying-parameter convergent-differential neural network for solving joint-angular-drift problems of redundant robot manipulators," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 2, pp. 679–689, Apr. 2018.

[27] Z. Zhang and L. Zheng, "A complex varying-parameter convergent-differential neural-network for solving online time-varying complex sylvester equation," *IEEE Trans. Cybern*, pp. 1–13, 2018.

[28] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[29] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[30] X. Li, X. Jia, G. Xun, and A. Zhang, "Improving eeg feature learning via synchronized facial video," in *Proc. IEEE Int. Conf. Big Data*, 2015, pp. 843–848.

[31] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[32] Y. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, Oct. 2018.

[33] Y. Ding, X. Hu, Z. Xia, Y. Liu, and D. Zhang, "Inter-brain EEG feature extraction and analysis for continuous implicit emotion tagging during video watching," *IEEE Trans. Affect. Comput.*, pp. 1–12, 2018.

[34] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2image: Converting brain signals into images," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1809–1817.

[35] M. Bilalpur, S. M. Kia, T. S. Chua, and R. Subramanian, "Discovering gender differences in facial emotion recognition via implicit behavioral cues," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 119–124.

[36] Y.-P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010.

[37] S. Dähne, F. Bießmann, F. C. Meinecke, J. Mehnert, S. Fazli, and K.-R. Müller, "Integration of multivariate data streams with bandpower signals," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1001–1013, Aug. 2013.

[38] F. Cong *et al.*, "Linking brain responses to naturalistic music through analysis of ongoing EEG and stimulus features," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1060–1069, Aug. 2013.

[39] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng.*., 2013, pp. 81–84.

[40] Y. Zhong and Z. Jianhua, "Cross-subject classification of mental fatigue by neurophysiological signals and ensemble deep belief networks," in *Proc. 36th Chin. Control Conf.*, 2017, pp. 10 966–10 971.

[41] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted Boltzmann machines," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 566–576, Jun. 2017.

[42] A. Gogna, A. Majumdar, and R. Ward, "Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2196–2205, Sep. 2017.
[43] A. Antoniades *et al.*, "Detection of interictal discharges with convolutional neural networks using discrete ordered multichannel intracranial EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 12, pp. 2285–2294, Dec. 2017.
[44] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
[45] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, Feb. 2018.
[46] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep feature learning for EEG recordings," 2015, ArXiv:1511.04306, 2015.
[47] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional network for EEG-based brain-computer interfaces," 2016, arXiv:1611.08024.
[48] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6809–6817.
[49] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1–9.
[50] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.
[51] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. Advances Neural Inf. Process. Syst.*, 2011, pp. 1017–1025.
[52] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.
[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.
[54] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

Jianmin Jiang received the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 1994.

From 1997 to 2001, he worked as a Full Professor of Computing with the University of Glamorgan, Wales, U.K. In 2002, he joined the University of Bradford, Bradford, U.K., as a Chair Professor of Digital Media and Director of the Digital Media & Systems Research Institute. He worked with the University of Surrey, Surrey, U.K., as a Full Professor during 2010–2014 and a Distinguished Chair Professor (1000-plan) with Tianjin University, Tianjin, China, during 2010–2013. He is currently a Distinguished Chair Professor and Director with the Research Institute for Future Media Computing at the College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China. He has authored approximately 400 refereed research papers. His research interests include image/video processing in the compressed domain, digital video coding, medical imaging, computer graphics, machine learning, and AI applications in digital media processing, retrieval, and analysis.

Dr. Jiang was a Chartered Engineer, Fellow of IEE, Fellow of RSA, member of EPSRC College in the U.K., and EU FP-6/7 evaluator.

Ahmed Fares received the Ph.D. degree from the Department of Computer Science and Engineering, Egypt-Japan University of Science and Technology (E-JUST), New Borg El Arab, Egypt, in 2015.

Currently, he is a Postdoctoral Researcher with the Research Institute for Future Media Computing at the College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China, and an Assistant Professor with the Department of Electrical Engineering and the Computer Engineering branch at the Faculty of Engineering at Shoubra, Benha University, Cairo, Egypt. His research interests include brain science, cognitive science, computational modeling, theoretical computer science, and machine learning.

Sheng-Hua Zhong received the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2013.

She worked as a Postdoctoral Research Associate with the Department of Psychological & Brain Sciences at The Johns Hopkins University, Baltimore, MD, USA, from 2013 to 2014. Currently, she is an Assistant Professor with the College of Computer Science & Software Engineering at Shenzhen University, Shenzhen, China. Her research interests include multimedia content analysis, cognitive science, psychological and brain science, and machine learning.