A Brain-Media Deep Framework Towards Seeing Imaginations Inside Brains

Jianmin Jiang, Ahmed Fares*, and Sheng-hua Zhong

***Abstract*—While current research on multimedia is essentially dealing with the information derived from our observations of the world, internal activities inside human brains, such as imaginations and memories of past events etc., could become a brand new concept of multimedia, for which we coin as "brain-media". In this paper, we pioneer this idea by directly applying natural images to stimulate human brains and then collect the corresponding electroencephalogram (EEG) sequences to drive a deep framework to learn and visualize the corresponding brain activities. By examining the relevance between the visualized image and the stimulation image, we are able to assess the performance of our proposed deep framework in terms of not only the quality of such visualization but also the feasibility of introducing the new concept of "brain-media". To ensure that our explorative research is meaningful, we introduce a dually conditioned learning mechanism in the proposed deep framework. One condition is analyzing EEG sequences through deep learning to extract a more compact and class-dependent brain features via exploiting those unique characteristics of human brains such as hemispheric lateralization and biological neurons myelination (neurons importance), and the other is to analyze the content of images via computing approaches and extract representative visual features to exploit artificial intelligence in assisting our automated analysis of brain activities and their visualizations. By combining the brain feature space with the associated visual feature space of those images that are candidates of the stimuli, we are able to generate a combined-conditional space to support the proposed dual-conditioned and lateralization-supported GAN framework. Extensive experiments carried out illustrate that our proposed deep framework significantly outperforms the existing relevant work, indicating that our proposed does provide a good potential for further research upon the introduced concept of "brain-media", a new member for the big family of multimedia. To encourage more research along this direction, we make our source codes publicly available for downloading at GitHub[1]**

***Index Terms*—EEG, Image Generation, Deep Learning, Brain Media, Bi-directional Computation, Variant LSTM.**

I. Introduction

MULTIMEDIA has been extensively researched over the past decades, in which all the forms of multimedia, such as texts, audios, images and videos etc. can be regarded as external, since all of them essentially record what we see rather than what we thought. Yet some elements of the thoughts, such as imaginations, aspirations, and emotional memories etc., could be visualized and reproduced into a new form of multimedia. For the convenience of presentation, we coin such a new form of multimedia, reflecting the internal world inside human brains, as "brain-media". As a matter of fact, studies on brain activities, especially via electroencephalograms (EEGs), have been researched across a number of areas, including neuroscience, brain science, psychology and computer science [1], [2], [3], [4]. For the past decades, research on understanding brain activities has been active through EEGs evoked by specifically designed stimuli for brain computer interfacing (BCI) [5], [6], [7], and studies in both psychology and neuroscience reveal that up to a dozen of special categories can be recognized by event-related potential (ERP) recorded via EEGs [8], [2]. Further, a range of machine learning models [9], [10], [11] have also been developed to address the problem of multimedia-evoked brain understanding through approaches of pattern recognition and classifications, and many improved results have been reported in the literature. In this paper, we push the existing EEG-based brain research a step further and promote such research towards the direction of introducing a new concept of multimedia, i.e. brain-media, and hence explore the possibility of enabling people to see what we thought rather than what we see. To turn such an ambitious notion into a feasible research direction, we propose a GAN-based deep

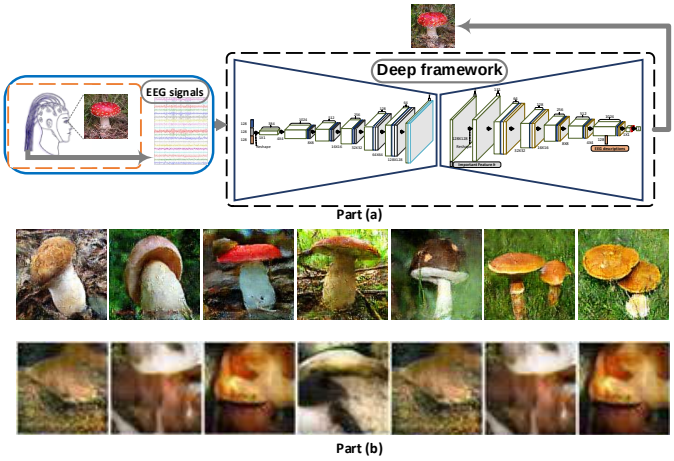

Fig. 1. Illustration of our research work on visualization of brain activities via the proposed deep framework (part-(a)) and the scenarios of our achieved experimental results (part (b)).

Jianmin Jiang, Ahmed Fares, and Sheng-hua Zhong are with the Research Institute for Future Media Computing, College of Computer Science & Software Engineering, Shenzhen University, China; Jianmin Jiang, jianmin.jiang@szu.edu.cn; Ahmed Fares, ahmed.fares@szu.edu.cn; Sheng-hua Zhong, csshzhong@szu.edu.cn.

Ahmed Fares, Corresponding author.

Jianmin Jiang and Sheng-hua Zhong are with the Guangdong Laboratory of Artificial Intelligence & Digital Economy (SZ), Shenzhen University, Shenzhen, China

Ahmed Fares is with the Department of Electrical Engineering, Computer Engineering branch, Faculty of Engineering at Shoubra, Benha University, Egypt; ahmed.fares@feng.bu.edu.eg

Manuscript received X X, X; revised X X, X.

[1]https://github.com/aneeg/LS-GAN

framework to visualize those brain activities evoked by natural images.

Fig. 1 illustrates the concept of our proposed research as well as some samples of the visualized results. As seen in part (a), the brain is evoked by a "bolete" image and its activated EEG sequences are fed into our proposed GAN-based deep framework. Via deep adversarial learning of the EEG sequences, an image is generated at the output to visualize the cognitive activities inside the human brain, which corresponds to its responses to the stimulating image at the input. By comparatively examining the relevance of the two images between the input stimuli and the output, we are able to evaluate the visualization performances of the proposed deep framework in terms of both the accuracy of the visualization and the quality of the visualization. Part (b) of Fig. 1 illustrates two rows of sample images generated as the visualization outputs. While the first row of samples are the "bolete" images generated by our proposed deep framework, the second row represents the samples generated by the existing research, brain2image GAN [12], which is used as the benchmark for evaluating our proposed deep framework. As seen, our generated output images have significantly better quality than that of the benchmark, indicating a high level of improvement by our proposed. Further experiments reported later in this paper also validate that our proposed achieves a much better accuracy, too, as supported by the corresponding classification results.

In addition to the ambitious concept of brain-media to enable us to see what we thought, our proposed research reported in this paper has twofold contributions, which can be highlighted as: (i) our proposed dually conditioned GAN exploits the interactions across both the brain domain and the visual domain to enhance the adversarial learning and hence achieve significantly improved results on both visualization quality and accuracy. While the brain domain provides essential support for our proposed deep framework to capture the thoughts-related activities, the visual domain captures the characteristics in visual content to help with high quality visualization of those brain thoughts. (ii) inspired by the phenomenon of hemispheric lateralization and the attention mechanism, we propose a new attention-gated LSTM to emphasize the differences between two hemispheres with low dimensions and measure the importance of different EEG channels. In this way, we are able to strengthen the capability of the proposed deep framework in brain representations and learning towards improved performances for its visualizations.

The rest of the paper is organized as follows. In section II, we present descriptions of relevant work that use deep learning models for EEG-based image visualization and classification. In Section III, we describe the details of our proposed deep framework for visualizing brain activities into images. In Section IV, we report our extensive experimental results and validate the superiority and effectiveness of our proposed framework, in comparison with the existing state-of-the-arts, and finally Section V provides concluding remarks and future work.

II. Related Work

Although the concept of brain-media has never been explored before, relevant research on EEG-based brain analysis has been extensive across a number of areas, including brain science, psychology, bio-engineering, and computer science, in which brain activities can be recorded using multiple techniques, including fMRI, EEG, and MEG, whose temporal and spatial resolutions have allowed computational methods to decode specific visual stimuli. Before the popularity of deep learning, Kaneshiro et al. [11] proposed a representational similarity based linear discriminant analysis framework to classify visually evoked EEG data according to twelve different object categories, and an accuracy of 28.87% was reported on their proposed dataset, ObjectCategory-EEG dataset. Since the advent of deep learning models, numerous new research attempts have been reported to leverage its strength in designing ambitious algorithms to achieve better understanding of brain activities and reconstructing a perceived visual stimulus via EEGs. Examples of such efforts can be illustrated by classification of brain activities via mining of EEGs. Kulasingham et al. [2] used deep belief networks (DBN) and deep automatic encoders to represent EEGs for detecting special patterns, and an average precision rate of 86.9% is reported for the deep belief network and 86.01% for the stacked autoencoder on their dataset. Yin and Zhang [13] proposed a single-channel EEG classification method with a deep belief network, decoding mental loads from EEGs, and an average classification accuracy of 71% was achieved based on the non-overlapped training and testing of EEG sequences. Lu et al. [14] proposed a frequential DBN (FDBN) for the purpose of classifying the motor imagery. The FDBN is based on three restricted Boltzmann machines (RBMs) stacked with a SoftMax regression, in which wavelet packet decomposition (WPD) and fast Fourier transform (FFT) are employed to obtain the frequency domain representation of the EEG signals. Stober et al. [15] used convolutional neural networks and an autoencoder to classify audio-evoked EEG recordings with an accuracy of 28% over 12 songs. Ogawa et al. [16] used a recurrent neural network (RNN) to simultaneously input video features and video viewers' EEG signals to achieve video classification based on user preferences. Spampinato et al. [17], used long short term memory (LSTM) network to learn an EEGs feature representation based on visual stimuli and constructed a mapping relationship from deep learned visual features to EEG feature representations. Finally, they utilized their proposed representation of EEG signals for classifying natural images. Compared with other existing approaches, these deep learning-based methods have accomplished outstanding classification results on their dataset, ImageNet-EEG.

Since the work on reconstruction and classification of EEG data reported by Gogna et al [18], research on visualizing EEGs becomes active. Schirrmeister et al. [19] reported the effects of convolutional neural networks (CNNs) on decoding and visualizing EEGs. They evaluated a large number of CNNs on an EEG decoding task, and demonstrated that advances from the field of deep learning, including exponential linear units and batch normalization, are crucial for achieving

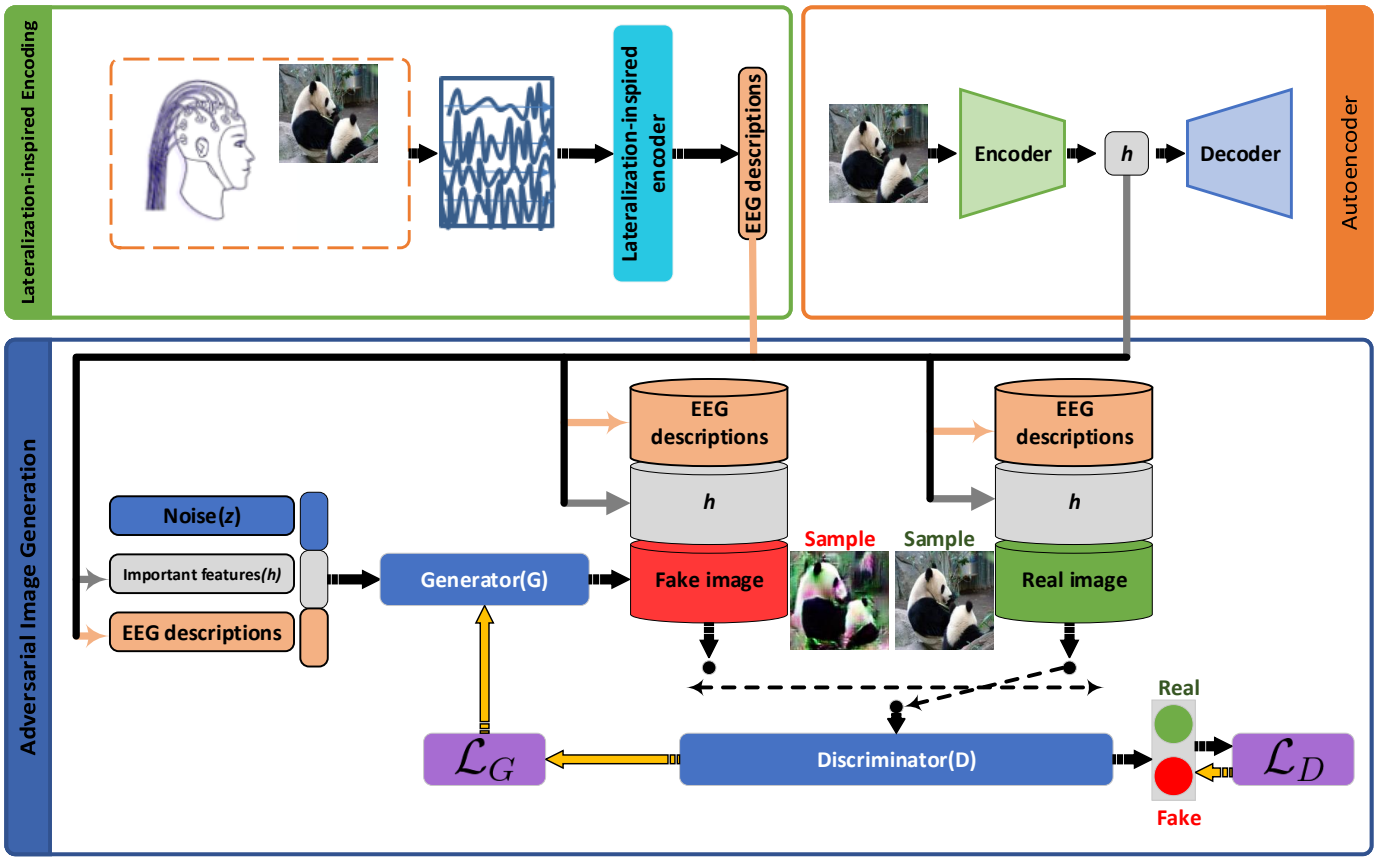

Fig. 2. Structural illustration of the proposed deep framework.

high decoding accuracies. Tirupattur et al. [20] proposed an EEG-based deep learning method, ThoughtViz, for visualizing human thoughts utilizing a GAN-based network. Kavasidis et al. [12] proposed a framework for generating the visual stimuli content information through EEG data. By using generative adversarial networks (GANs) [21] and variable-valued autoencoder (VAE) [22], they found that EEG data contain patterns related to visual content, and the content can be used to generate images that are semantically consistent with the input visual stimuli. Palazzo et al. [23] used conditional GAN-based framework [12] to generate visual stimuli through EEG data.

While these methods have demonstrated the capability of using deep learning for brain activity visualization and classification, it suffers from the flowing disadvantages: (i) the original EEG data or the extracted time-frequency features based on signal analysis algorithms are often used as the input only, and some characteristics of human brains have not been seriously considered, making the existing work on visualization of EEGs less indicative of actual brain activities; (ii) the importance of channel-based spatial information has not been exploited jointly with the information from the brain side, such as hemispheric lateralization, and hence the spatial and dynamic correlations embedded inside EEG sequences are relatively ignored; and finally (iii) the state-of-the-art Inception score (IS) and the classification accuracy achieved by Kavasidis et al. and Spampinato et al. [17] are only 5.07 and 82.9%, respectively, leaving a significant scope for further research and improvement.

To rectify the above weakness and move the existing work closer to the feasible exploration of the new multimedia concept: brain-media, we need to resolve two fundamental issues, which can be highlighted as: (i) significantly improve the accuracy of detecting those elements that can be visualized as brain-media out of the brain thoughts corresponding to stimulating images; (ii) significantly improve the quality of such visualizations and thus the visualized images can be enjoyed as any other natural images we generally encounter in the existing multimedia. To this end, we introduce a dually-conditioned and lateralization-supported GAN framework, where the brain feature space is combined with representation learning of visual feature spaces to provide a further assistance for deep learning of brain EEG sequences and improving the visualization performances. To achieve a seamless integration with EEG descriptions of brain cognitive responses to the external stimulation via natural images, we add a new regional attention gate into the existing LSTM to exploit the hemispheric lateralization [24] and produce the first condition to support the proposed deep learning framework. Hemispheric lateralization refers to the tendency for some neural functions or cognitive processes to be specialized to the right or the left hemispheres of human brains [25]. In addition, the attention mechanism, which allows a deep network to pay attention to only part of the input information, becomes one of the most powerful and

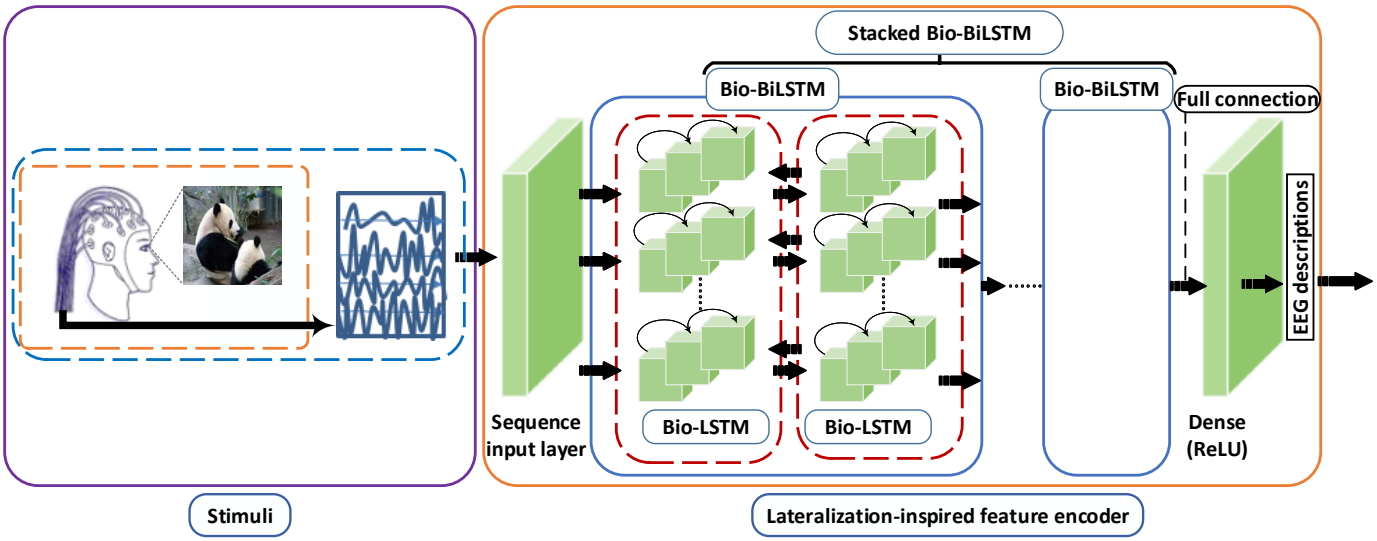

Fig. 3. Structural illustration of the proposed bio-inspired deep encoder.

influential ideas in deep learning [26], [27], [28], [29], [30]. As EEGs are channel-based temporal-spatial signal sequences, some parts of human brains are more deeply involved than others, leading to the oscillations in EEG signals and hence creating further spaces for improvement. To the best of our knowledge, no research has been attempted to integrate jointly the hemispheric lateralization and the attention mechanism into a gated structure of the recurrent deep learning model to extract the region-level information from brain signals.

III. The Proposed Brain-Media Deep Framework

Fig. 2 illustrates an overview of our proposed deep framework for visualizing the brain-media elements of the human thoughts corresponding to the natural image stimulation at the input, from which it can be seen that we use a lateralization-inspired LSTM to extract EEG descriptions and generate the first condition (top-left of Fig. 2), and we use an auto-encoder to learn data representations and extract visual features across all the candidate images for stimuli (top-right of Fig. 2) to generate the second condition.

In general, GAN consists of two networks, including generator (G) and discriminator (D). While the generator tries to create a sample from a random noise input $(\mathbf{z})$, the discriminator checks whether the generator is actually creating a fake sample or real sample. As the generator is expected to capture the overall training data distributions and generate realistic-looking samples, the discriminator would be regularly uncertain of whether its inputs are real or fake images. To visualize the image-evoked brain activities and ensure that such automated visualization can overcome the ambiguity and variation of the stimuli images brought to the brain activations, we need to maximize the potentials from both the brain side and the content side of natural images. To this end, we propose to generate image samples by both the random noise vector and the combined description vector, integrated from the two conditions as shown in Fig. 2, in order to enable the proposed brain-media deep framework to achieve the best possible visualization of the captured brain activities in terms of both quality and accuracy.

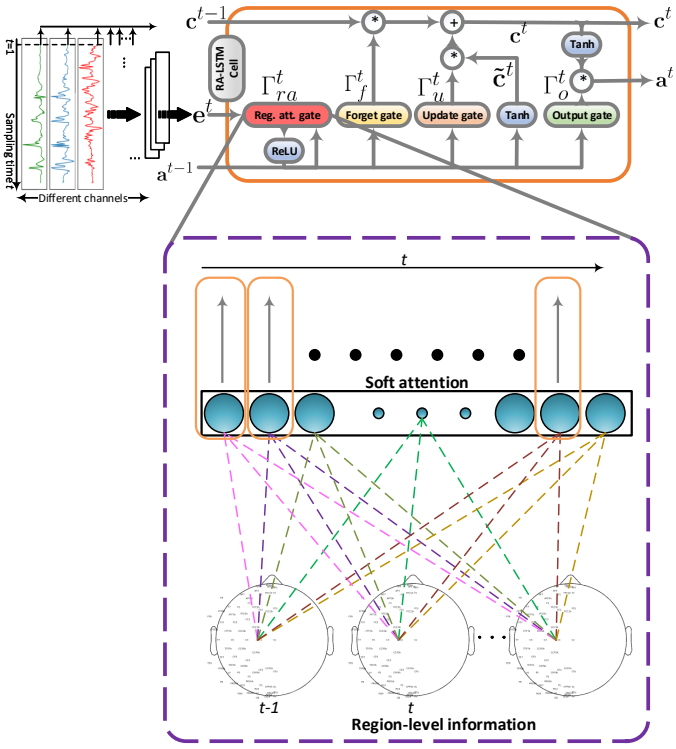

Fig. 4. Structural illustration of the proposed attention-gated LSTM cell.

To produce the first condition and exploit the EEG description of brain activities in the process of adversarial learning, we design a stack of n lateralization-inspired and bi-directional LSTM layers as shown in Fig. 3. Given the input $\mathbf{e}$ from all channels at time t, specifically, an additional gate, referred to as regional attention gate, is created to work together with the existing three gates, and hence their state values , i.e. the regional attention gate Γ_{ra}^t, the update gate Γ_u^t, the forget gate Γ_f^t, and the output gate Γ_o^t, which are represented by

colorful boxes in the attention-gated LSTM cell in Fig. 4, can be calculated from the raw EEG brain signals $\mathbf{E} = [\mathbf{e}_i]_{i=1}^{l_{ch}}$, where $i \in [1,\ l_{ch} = 128]$ is the index for EEG channels, l_{ch} is the number of EEG channels, and the previous layer output $\mathbf{a}^{t-1}$ is determined according to the following equation:

$$\begin{pmatrix} \Gamma_{ra}^t \\ \Gamma_f^t \\ \Gamma_u^t \\ \Gamma_o^t \end{pmatrix} = g \begin{pmatrix} \mathbf{W}_{ra} & \mathbf{U}_a & 0 \\ 0 & \mathbf{U}_f & \mathbf{W}_f \\ 0 & \mathbf{U}_u & \mathbf{W}_u \\ 0 & \mathbf{U}_o & \mathbf{W}_o \end{pmatrix} \begin{pmatrix} \left[\left(\mathbf{E}_{[l],j}^t - \mathbf{E}_{[r],j}^t\right)\ \mathbf{E}_{[m]}^t{}^{\mathrm{T}}\right] \\ \mathbf{a}^{t-1} \\ \Gamma_{ra}^t \end{pmatrix} + g \begin{pmatrix} \mathbf{b}_{ra} \\ \mathbf{b}_f \\ \mathbf{b}_u \\ \mathbf{b}_o \end{pmatrix} \tag{1}$$

where, for $k \in \{ra, f, u, o\}$, $\mathbf{W}_k$ is the weight matrix mapping the layer input to the four gates, $\mathbf{U}_k$ is the weight matrix connecting the previous cell output state to the four gates, and $\mathbf{b}_k$ is the bias vector. The function $g(.)$ is designed as ReLU activation function for Γ_{ra}^t and element-wise sigmoid for Γ_f^t, Γ_u^t, and Γ_o^t, respectively. To achieve the desired lateralization effect, the state of the regional attention gate Γ_{ra}^t is fed through the three gates. To achieve the desired lateralization effect, Γ_{ra} splits the EEG data into three groups, including the left hemisphere, the right hemisphere, and the central part. By denoting the left hemisphere group, the right hemisphere group, and the central group as, $\mathbf{E}_{[l]}$, $\mathbf{E}_{[r]}$, and $\mathbf{E}_{[m]}$, respectively, each channel $\mathbf{e}_i$ can be linked to one group based on its corresponding electrode physical location. In addition, each channel in the left hemisphere group has a corresponding channel in the right hemisphere group. Γ_{ra} combines the difference, $(\mathbf{E}_{[l],j}^t - \mathbf{E}_{[r],j}^t)$, and the central group, $\mathbf{E}_{[m]}$, into one variable, and then passes it to the attention part as an input, where $j \in [1,\ l_g]$ is the index for the left hemisphere, the right hemisphere, and l_g is the number of channels linked to the left hemisphere or the right hemisphere. To optimize the process of adding the regional attention-driven mechanism, we propose a soft regional attention gate, where the input EEG signals of different channels are fully connected with the nodes in the gate. As a result, the size of $\mathbf{W}_{ra}$ depends on the number of channels and the number of nodes in the regional attention gate. Based on the results of (1), the cell output state $\mathbf{c}^t$ and the layer output $\mathbf{a}^t$ (both the forward and the backward outputs) can be calculated from the state of the regional attention gate Γ_{ra}^t and the previous layer output $\mathbf{a}^{t-1}$, details of which are given below:

$$\mathbf{c}^t = \Gamma_f^t * \mathbf{c}^{t-1} + \Gamma_u^t * \overbrace{(\tanh(\mathbf{U}_c \mathbf{a}^{t-1} + \mathbf{W}_c \Gamma_{ra}^t + \mathbf{b}_c))}^{\text{Candidate for replacing the memory cell}} \tag{2}$$

$$\mathbf{a}^t = \Gamma_o^t * \tanh(\mathbf{c}^t) \tag{3}$$

where $\mathbf{W}_c$ is the weight matrix, mapping the layer input to the candidate for replacing the memory cell. While $\mathbf{U}_c$ is the weight matrix connecting the previous cell output state to the candidate for replacing the memory cell, $\mathbf{b}_c$ is a bias vector regulating the balance of strength between ($\mathbf{U}_c$ and ($\mathbf{W}_c$, and the function $\tanh(.)$ indicates a hyperbolic tangent. When other layers are present, the output of the first layer is provided as an input to the second layer and so on. The final output of the deepest LSTM layer is a vector of all outputs, represented by $\mathbf{Y} = [\mathbf{y}^t]_{t=1}^{l_s}$. At each time of iteration t, $\mathbf{y}^t$ can be calculated according to the standard LSTM equation [31]. For our proposed brain-media deep framework, however, only the last element of the output vector, $\mathbf{y}^{l_s}$, is taking into account as a candidate to represent the first condition.

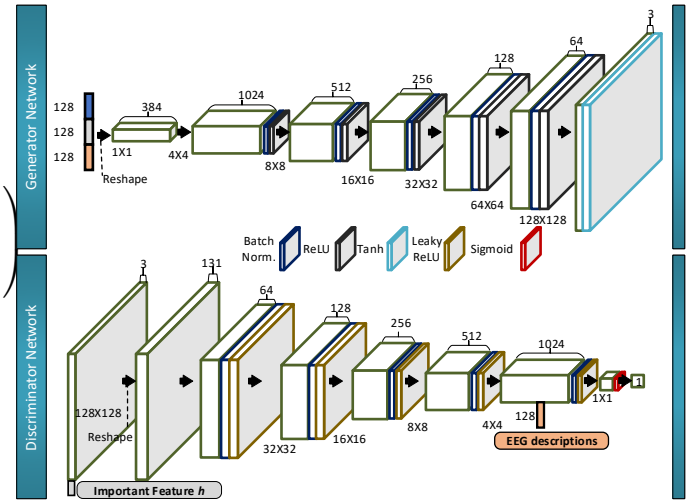

Fig. 5. Structural illustration of the proposed dually-conditioned and lateralization supported GAN.

For visualization, the generator network $G(\mathbf{z}|\mathbf{y}, \mathbf{h})$ is trained in a conditional GAN framework to map the random inputs from a $p_z(\mathbf{z})$ noise distribution and the combined-conditional vector, including EEG descriptions ($\mathbf{y}$) and the visual features ($\mathbf{h}$), to a target image distribution $p_{data}(\mathbf{x})$ as seen in Fig. 2. We train both the generator and the discriminator at the same time in a minimax gaming environment. While the generator attempts to maximize the probability of making the discriminator mistake its inputs $p_G(\mathbf{z}|\mathbf{y}, \mathbf{h})$ as real, the discriminator attempts to maximize the probability of associating the correct labels, i.e. real samples to $p_{data}(\mathbf{x})$ and fake samples to $p_G(\mathbf{z}|\mathbf{y}, \mathbf{h})$. The overall objective function $\mathbf{V}(D, G)$ can be calculated according to the following equation:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x} \in p_{data}} [\log D(\mathbf{x}|\mathbf{y}, \mathbf{h})] + \mathbb{E}_{\mathbf{z} \in p_{\mathbf{z}}} [\log (1 - D(G(\mathbf{z}|\mathbf{y}, \mathbf{h})|\mathbf{y}, \mathbf{h}))] \tag{4}$$

The discriminator loss function $\mathcal{L}_D$ and the generator loss function $\mathcal{L}_G$ are implemented using a hinge loss, where we modify the generator loss function by adding the constrictive loss to the adversarial loss. To justify our choice, we have also investigated using Wasserstein GAN loss via empirical studies. While the inception score (IS) achieved by WGAN loss is 6.59, the IS value becomes 6.64 by using hinge loss. Further experiments reveal that, by using either the hinge loss or the WGAN loss, our proposed framework still outperforms the compared benchmarks.

Precisely, an image generated by a generator network is passed as an input to an accompanied constructive loss that evaluates the discrepancy between the generation results and

TABLE I
THE EXPERIMENTAL PARAMETERS

Parameters/Dataset	ImageNet-EEG	ObjectCategory-EEG
No. of classes	40	6
No. of stimuli per class	50	12
Total No. of stimuli	2000	72
No. of trails per subject per stimuli	1	72
No. of subjects	6	10
Time for each stimuli	500ms	500ms

the ground truth. Details of the discriminator and the generator loss functions, $\mathcal{L}_D$ and $\mathcal{L}_G$, can be described as follows

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}[\min(0, -1 + D(\mathbf{x}|\mathbf{y}, \mathbf{h}))] \\ &- \mathbb{E}[\min(0, -1 - D(G(\mathbf{z})|\mathbf{y}, \mathbf{h}))] \end{aligned} \quad (5)$$

$$\mathcal{L}_G = -\alpha\ \mathbb{E}[D(\mathbf{z}|\mathbf{y}, \mathbf{h})] + \beta\ \ell_1(p_G(\mathbf{z}|\mathbf{y}, \mathbf{h}),\ p_{data}(\mathbf{x})) \quad (6)$$

where α and β are the two weighting coefficients balancing the contribution of the adversarial and constrictive losses.

Fig.5 is a structural illustration of the proposed adversarial image generation network, including the generator G and discriminator D , which are implemented as two convolutional neural networks inspired by the DCGAN [32]. As seen in the generator network G, the combined-conditional vector $(\mathbf{y}, \mathbf{h})$ is concatenated with the random noise vector $\mathbf{z}$ and a series of deconvolutions are designed to upsample the concatenated vector to an output image. For the discriminator network D, it starts by receiving an image, either real or generated image, concatenated with the condition vector $\mathbf{h}$ associated with the input image. As opposed to the G network, the D network performs a series of convolutions, each of which reduces the size of the feature map spatial dimensions, and then appends the conditional vector $\mathbf{y}$, associated with the input image, to the last convolutional layer. Finally, the discriminator also calculates the output probabilities.

IV. EVALUATIONS AND EXPERIMENTAL RESULT ANALYSIS

To evaluate our proposed brain-media deep framework, we have carried out extensive experiments and assessed the brain-media visualization performances in terms of both accuracy and quality. For visualization accuracy, we apply the attention-gated LSTM encoder, part of our proposed deep framework, to classify the EEG descriptions of brain signals into a category of candidate images for stimuli and see if the classified category remains the same as that of the input image or not. The precision rate of such a classification is then used to measure the accuracy of the proposed visualization. To evaluate the quality of the proposed brain-media visualization, we primarily inspect the generated output images and assess their quality on subjective perception basis in the same way as that for the existing multimedia (images), although some quantified testing is also carried out via Inception score (IS) measurements.

A. Datasets

To evaluate our proposed brain-media visualization and analyze the achieved experimental results on a comparative basis against the existing efforts, we adopt two standard datasets for EEG-based brain activity descriptions: ImageNet-EEG [17] and ObjectCategory-EEG [11], in order to reduce the risk of overfitting to a particular dataset and limiting the generality of our research.

ImageNet-EEG dataset is a publicly available EEG dataset for brain activity viualization and classification prepared by Spampinato et al. [17], which is gathered utilizing a 128-channel cap with active, low-impedance electrodes (acti-CAP 128Ch). It incorporates the EEG signals of six subjects, one female and five male, produced by requesting them to look at the visual stimuli, which are images chosen from a subset of ImageNet (ILSVRC) [33]. It incorporates 40 classes, including "dog", "cat", "butterfly", "sorrel", "capuchin", "elephant", "panda", "fish", "airliner", "broom", "canoe", "phone", "mug", "convertible", "computer", "watch", "guitar", "locomotive", "espresso", "chair", "golf", "piano", "iron", "jack-o′-lantern", "mailbag", "missile", "mitten", "bike", "tent", "pajama", "parachute", "pool", "radio", "camera", "gun", "shoe", "banana", "pizza", "daisy" and "bolete" (fungus), and each class has 50 images. During the subjective experiment, each image has appeared on the computer screen for 500 ms. Table I summarizes the experimental parameters. The sampling frequency and data resolutions are set to 1kHz and 16 bits, respectively.

The EEG data in ImageNet-EEG has been filtered by a notch filter (49-51 Hz) and a second-order band-pass Butterworth filter (low cut-off frequency 14 Hz, high cut-off frequency 71 Hz). Therefore, the recorded signals only included the Beta (15-31 Hz) and Gamma (32-70 Hz) rhythm bands. As known, these bands have information about the cognitive process and perceptions [34].

The second EEG dataset we used to evaluate the effectiveness of our proposed brain-media visualization is an open-access dataset compiled at Stanford University by Kaneshiro et al. [11]. ObjectCategory-EEG is collected using unshielded 128-channel EGI HCGSN 110 nets, it includes the EEG signals of 10 subjects, aged 21 to 57 years (3 female), produced by asking them to look at the visual stimuli, which are images selected from a subset of the 92-image set used in other representational similarity analysis (RSA) studies [35], [36]. ObjectCategory-EEG consists of 6 classes, including "human body (HB)", "human face (HF)", "animal body (AB)", "animal face (AF)", "fruit vegetable (FV)", and "inanimate object (IO)", and each class has 12 images. For EEG signal collection, each image was shown on the computer screen 12 times at a random order for a total of 864 trials per recording. Each trial consisted of a single image shown on screen for 500 ms, and the subject completed two experimental sessions, each of which contains three blocks of 864 trials, resulting in a total of 5184 trials per subject. Table I summarizes the experimental parameters.

The EEG data in ObjectCategory-EEG has been filtered by a high-pass fourth-order Butterworth filter for removing frequency content below 1 Hz and a low-pass eighth-order Chebyshev Type I filter for removing frequencies above 25 Hz. The sampling frequency was set to 1kHz with a range of 24 bits. Across the recordings in the ObjectCategory-EEG

dataset, preprocessing is applied to achieve three effects: (i) interpolation of approximately five channels; (ii) removal of 0-2 trials for every image; and (iii) removal of four independent components.

B. Experimental Settings and Training Details

As shown in Fig. 2, the size of all layers in the attention-gated LSTM is set to 68, including those subsequent non-linear layers, and there are two layers in the stacked BiLSTM ($n = 2$). The iteration limit is set to 2500, and the batch size is set to 440. For the autoencoder, the encoder network consists of four deconvolutional layers, which takes an input image and return the image representations, i.e. the visual features. On the other hand, the decoder network consists of five convolutional layers, which takes the image representation as the input and tries to return the same image as an output. The number of epochs is set to 200, and the batch size is set to 128.

For those networks inside the GAN, as seen in Fig. 5, the generator network takes a 384-concatenated-dimensional vector as the input, including 128-dimensional random noise vector z and a combined-conditional vector consists of 128-dimensional visual features h and 128-dimensional propped encoder output y. It reshapes this input vector to a 4-dimensional vector and then feeds it to a sequence of five layers, each consists of three operations, including deconvolutions, Batch Normalization, and ReLU operation. This sequence of operations doubles the spatial dimensions of the input vector while halves its number of channels before the last one, which outputs a $128 \times 128 \times 3$ RGB colored image squashed between values of -1 and 1 through the `tanh` function. On the other hand, the discriminator network takes a concatenated input which consists of $128 \times 128 \times 3$ images and their associated 128-dimensional autoencoder output h. Similarly, the input is also reshaped into a 4-dimensional vector and then fed into a sequence of five layers, each of which consists of three operations, including convolutions, Batch Normalization, and Leaky ReLU operation. This sequence of operations halves the spatial dimensions of the input while doubles its number of channels, and the output of the last convolutional layer is concatenated with its associated 128-dimensional EEG descriptions y. The last layer is flattened and then fed into a single sigmoid output.

Training for deep learning based generative adversarial models is a challenging problem for two reasons: i) balancing the generator and discriminator; ii) overfitting due to the size of the dataset. For the first reason, we have investigated the two-timescale update rule (TTUR) technique [37] for unbalancing the learning rate between the generator and the discriminator updates. As reported by Goodfellow et al [38], TTUR is more effective than other approaches, such as spectral normalization, in stabling the training of GANs, and thus we propose to use the TTUR technique that provides different learning rates for the discriminator and the generator, in order to compensate for the slow learning rate of the discriminator. While we appreciate the point that there may exist many other state of the arts, our choice achieves additional advantages that: (a) fair comparisons with the existing benchmarks can be easily implemented by disabling the self-attention layer; and (b) SAGAN is essentially originated from DCGAN, which covers most of the state-of-the-art research on GANs, providing wider comparability and compatibility with the main stream of research on GANs. Specifically, we set the discriminator learning rate as 0.0004, and the generator learning rate at 0.0001, making it possible to utilize fewer generator steps for every single discriminator step. For the second reason, we use the largest dataset, ImageNet-EEG, to provide sufficient space for us to investigate the overfitting problem. The number of images with the associated EEG recordings, however, is very low with 50 recording per class, making either the generator or the discriminator to overfit if we directly train these two networks on it. As a result, we train the proposed GAN in two phases. In the first phase, we train the proposed deep framework using only images from ImageNet [33] with no EEGs for 100 epochs. During this phase, the attention-gated LSTM conditional vector is set to zero. Then in the second phase, we re-trained the models on the images with EEGs for 50 more epochs. During the training process, data is augmented by resizing images at 143×134 pixels, extracting random 128×128, and flipping images horizontally with a chance of 50%. Our deep framework is implemented on a Tesla® P100 GPU.

TABLE II
COMPARATIVE CLASSIFICATION RESULTS BETWEEN OUR PROPOSED ENCODER AND THE EXISTING BENCHMARKS, INCLUDING RNN-BASED METHOD, SIAMESE NETWORK, MULTIMODEL NETWORK, COGNINET, AND RS-LDA.

Models	Accuracy
Proposed encoder	99.1%
RNN-based model [17]	82.9%
Siamese network [40]	93.7%
Multimodal network [41]	94.1%
CogniNet [42]	89.6%
RS-LDA [11]	13.0%

For benchmarking purposes, the proposed brain-media visualization framework is compared with the EEG-based image generation methods [12], [23], which are the most recent deep learning methods conditioned by brain signals on the ImageNet-EEG dataset. In this research, we utilize the Inception score (IS) [39] that is a popular metric for judging the GANs output images by analyzing two conditions simultaneously, i.e. (i) the images should have varieties of meaningful content, and (ii) the image quality should be perceptually high, the same as those natural images. If both conditions are true, the floating-point score will be large. If either or both are false, the floating-point score will be small.

C. Assessment of Visualization Accuracy

To evaluate the visualization accuracy of our proposed deep framework, we carry out experiments to test the EEG-based classification performances of our proposed framework in comparison with the existing work. Such a design is based on the fact that our proposed deep framework relies on the EEG description and its corresponding deep understanding of the brain responses to the stimulation image to visualize what is happening inside human brains, and hence the clas-

TABLE III
FURTHER EXPERIMENTAL COMPARISONS BETWEEN OUR PROPOSED ENCODER AND RS-LDA.

Model/No. of classes	6-classes	2-classes Faces vs objects
Proposed encoder	61.10%	89.06%
RS-LDA [11]	40.68%	81.06%

sification results achieved by the deep framework indicates which category the visualized image belongs to, providing an indirect but logic measurement of the visualization accuracy. We use two publicly available datasets, ImageNet-EEG [17] and ObjectCategory-EEG [11], in order to reduce the risk of overfitting to any particular dataset and limiting the generality of our research. While the ImageNet-EEG received critiques recently from Purdue scholars [43], we feel that inclusion of this data set in our experiments still remains worthwhile until the critique is positively responded by the original authors and validated by the research communities. Table I summarizes the experimental parameters, and Table II summarizes the experimental results in terms of the classification precisions for our proposed attention-gated LSTM encoder and 5 benchmarks representing the existing state of the arts, including the RNN-based method [17], Siamese network [40], multimodal network [41], and CogniNet [42], and the RS-LDA method [11]. As seen, while the precision rate accomplished by our proposed encoder network is 98.4%, the RNN-based method, Siamese network, multimodal network, CogniNet, and RS-LDA compared are 82.9%, 93.7%, 94.1%, 89.6% and 13.0%, respectively. As a result, such significant improvement on classification precision achieved by our proposed deep encoder validates our contribution in developing the attention-gated LSTM and exploitation of lateralization for the encoding model design.

To quantify the contribution of integrating the regional attention gate in our proposed attention-gated LSTM deep encoder, we further carried out experiments to explore the effectiveness of different configurations made by proposed encoder. While the precision rate accomplished by the proposed encoder network without the regional attention gate is 95.3%, the precision rate accomplished our proposed attention-gated LSTM encoder is 99.1%.

To enable comparative assessment and result analysis against the existing benchmarks, we carried out further experiments to validate the effectiveness of our encoder network for EEG-based classification on ObjectCategory-EEG dataset [11]. In this experiment, we have used the same experiment set up as the existing work [11].

Table III summarizes the experimental results in terms of the classification precisions for our proposed encoder and the RS-LDA method [11]. As seen, while the precision rates accomplished by our proposed encoder are 61.10% and 89.06%, the RS-LDA compared are 40.68% and 81.06% on six-classes and two-classes, respectively. From these results, we claim: (i) the generality of our proposed attention-gated LSTM encoder is validated; (ii) the generalization capability of our proposed encoder is better than that of RS-LDA [11], supported by the better results achieved upon both data sets, ObjectCategory-EEG and ImageNet-EEG; (iii) the improvement achieved by our proposed has almost the same ratio upon both data sets, which is about 20%.

	HB	HF	AB	AF	FV	IO
HB	50.02	6.60	6.60	15.60	6.60	14.60
HF	0.88	95.49	0.88	1.88	0.00	0.88
AB	4.87	4.87	60.63	14.87	2.87	11.87
AF	9.77	7.77	7.77	62.15	6.77	5.77
FV	10.01	7.01	13.01	7.01	43.96	19.01
IO	9.98	4.98	5.98	4.98	14.98	59.11

Fig. 6. Illustration of confusion matrix for our proposed encoder.

TABLE IV
COMPARATIVE ASSESSMENT OF THE INCEPTION SCORE BETWEEN THE PROPOSED DEEP FRAMEWORK AND THE EXISTING BENCHMARKS.

Model	Inception score
proposed	6.64
EEG-based GAN [23]	5.07
brain2image GAN [12]	5.07
brain2image VAE [12]	4.49

For the convenience of further analysis and comparative investigation, Fig.6 presents the confusion matrix of each category for ObjectCategory-EEG. As the matrix diagonal represents the highest value in each row of the confusion matrix, labels predicted by the proposed encoder were most often the correct class labels. As seen, while the classification accuracy of the "human face (HF)" category is better than others, with 95.49% of trials being labeled correctly, the lower right portion of the confusion matrix, such as the "fruit vegetable (FV)" and "inanimate object (IO)", show notable confusion. These findings are consistent with what is reported in the existing work [11].

D. Testing on Visualization Quality

To evaluate the quality of our proposed brain-media visualization, we primarily apply our deep framework to ImageNet-EEG dataset and compare our visualized output images with those produced by the existing efforts [12], [23] with the same experimental settings.

To quantify the contributions of our deep framework to the quality of brain-media visualizations, we also computed the Inception Score (IS) on 50000 generated sample images, i.e, each class generates a sample of 1250 images. Table IV summarizes the experimental results in terms of the Inception scores for our proposed deep framework and the existing state-of-the-arts, including the EEG-based GAN [23], brain2image GAN [12], and brain2image VAE [12]. As seen, while the Inception Score achieved by our proposed deep framework is 5.89, the Inception score achieved by both EEG-based GAN and brain2image GAN is 5.07, and the Inception score achieved by brain2image VAE is 4.49.

To quantify the contribution of our deep framework, a further analysis is conducted. In previous work [12], [23], high-quality results for three of the ImageNet-EEG visual classes,

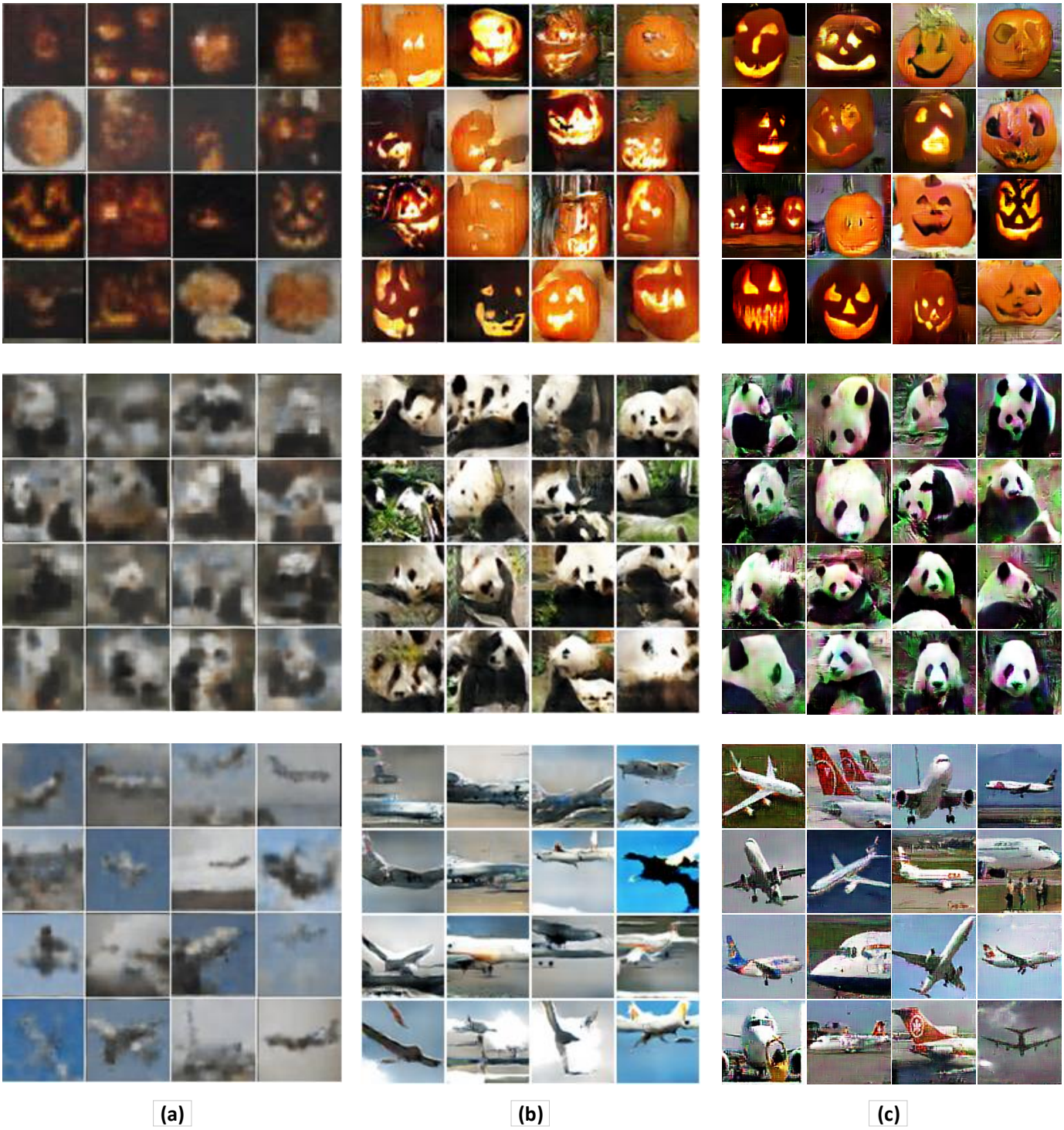

Fig. 7. Illustration of visualization samples for the classes of "jack-o′-lantern", "panda", and "airliner" achieved by: (a) brain2image VAE, (b) brain2image GAN, and (c) our proposed brain-media deep framework.

including "jack-o′-lantern", "panda", and "airliner", and low-quality results for other three of the ImageNet-EEG visual classes, including "banana", "capuchin", and "bolete" , are reported and illustrated for visual inspections and subjective assessments. We follow the same strategy and demonstrate two sets of generated samples by our proposed and the 3 existing benchmarks in Fig. 7 and Fig. 8, in which the two benchmarks, brain2image GAN and EEG-based GAN, are from the same authors and the samples generated by these two benchmarks are the same too. Consequently, the following observations can be made: (i) the quality of the visual content across the compared models is different; (ii) our proposed deep framework is able to translate the EEG descriptions into a meaningful and class-dependent images than other compared models; and (iii) the visual quality of our generated images on all compared classes are better that of the compared three

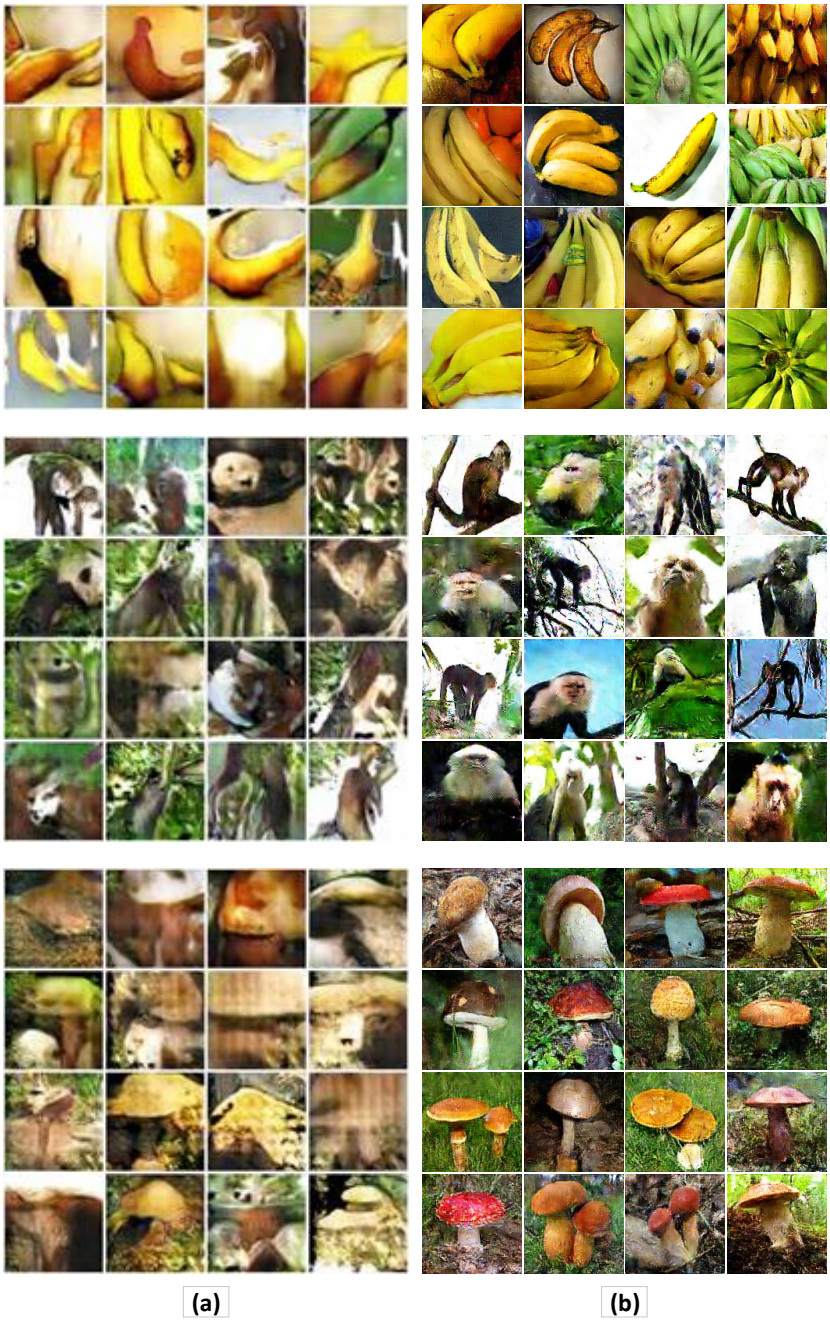

Fig. 8. Illustration of the visualization samples for the classes of "banana", "capuchin", and "bolete" achieved by: (a) brain2image GAN and EEG-based GAN, and (b) Our proposed brain-media deep framework.

benchmarks. While the first and the third observations validate our contribution to the quality improvement upon the proposed GAN model in generating images, the second observation validates our contribution in introducing the new concept of "brain-media" and illustrates the potential for developing it into feasible research directions, where semantic elements of the brain thoughts can be captured and visualized into enjoyable multimedia content by EEG-based deep learning models.

As indicated by the existing research across areas of brain science, psychology, and neural computations etc, finally, the primary means for interfacing with human brains at present are two approaches, via EEG or fMRI [44], [45]. In other words, brain responses to the external stimuli are primarily represented either by EEG sequences or by fMRIs. To this end, our proposed model is essentially developed to deeply learn and extract, out of those EEG sequences, the brain

TABLE V
COMPARATIVE ASSESSMENT OF THE PROPOSED FRAMEWORK UNDER DIFFERENT CONFIGURATIONS.

Training of the proposed framework	Testing of the proposed framework	Inception Score
Dually conditioned	Dually conditioned	6.64
Dually conditioned	Brain feature space only	6.39
Brain feature space only	Brain feature space only	6.34

responding activities to the external stimulations of the input image. When we classify EEGs via our proposed attention-gated LSTM encoder, for example, we are essentially trying to interpret which category of the input image the brain activity is responding to. When we visualize the EEGs via the proposed GAN model, equally, we are trying to visualize the brain activities responding to the input stimuli image. Consequently, the visualized image becomes a form of our introduced brain-media as long as: (a) its quality is good enough to be enjoyed in the same way as that of any natural image, and (b) its content is meaningful to human understandings and perceptions. To support the statement further, we carry out additional ablation studies and report the results in Table V. As seen from the results of our ablation studies, EEG sequences provide primary support for learning and analyzing brain activities responding to natural image stimulations. While the additional condition added from the visual feature space only provides marginal support for brain activity analysis as suggested by the small difference in IS scores, the visualization results indicate that the additional condition plays a significant role in improving the quality of the generated output image.

V. Conclusions

By integrating the brain feature space and visual feature space together, we have described in this paper a novel dual-conditioned and GAN-based deep framework for brain-media visualization, exploring the new concept of brain-media and providing a good potential to turn this new concept into a new member in the family of multimedia. Our proposed framework provides an improved solution for the problem that, given brain activities stimulated by an image, we should be able to learn more compact and class-dependent descriptions of the EEG signals and visualize what is responded by human brains, including those parts of brain activities containing semantic concepts such as objects and scenes etc. which can be visualized into meaningful brain-media. As the sensitivity level across different locations in generating EEG signals remains different, we introduce a regional attention gate into the existing LSTM and hence enable the regional attention gate to extract the region-level information to preserve and emphasize the hemispheric lateralization for neural functions or cognitive processes of human brains. In addition, the added regional attention gate also measures and seizes the importance of different EEG channels via the integrated channel-level attention mechanism, and hence drive the proposed brain-media deep framework to capture the dynamic correlations hidden in the EEG sequences. Extensive experiments on ObjectCategory-EEG and ImageNet-EEG, the most challenging EEG dataset publicly available for brain activity analysis, validate that our framework outperforms the existing state-of-the-arts under various contexts and experimental set ups. Further, our research has produced substantial evidences to support that the information captured straightforwardly from human brains has the potential to: (i) enable the developed machine learning models to make better and more human-like understandings of the cognitive process inside human brains; (ii) convey vision-related information towards multimedia description of brain responses to the external stimulations; and (iii) reconstruct the brain-perceived multimedia content via EEG representations and their deep learnings.

While the concept of "brain-media" we introduced in this paper has the potential of enabling computers to understand and interpret brain activities, it is not the same as "reading human minds or thoughts". Essentially, we are trying to capture those brain activities and thoughts that have meaningful semantics and can be visualized into enjoyable multimedia content, i.e. images or videos. In other words, not all human thoughts can be turned into brain-media, for which one typical example of meaningful semantics is those dreams that contain events and scenes. Due to the fact that human brain activities are enormously complicated, however, current research has to be limited to the environment that brain activities are stimulated via external images or graphics patterns in order to make the process manageable. It can be envisaged that, in the future, research upon the new concept of brain-media can be carried out in such a way that human brains can be monitored by computers to capture those meaningful pieces and visualized into brain-media without any external stimuli. As a result, the current research under external stimuli can be viewed as providing guidelines for the above objective similar to the scenario of training deep learning frameworks.

Finally, a range of further research can be identified to push forward the concept of brain-media and moving us to the stage that we are able to see the imaginations inside human brains, examples of which can be summarized as: (i) continuously monitoring human brains without any external stimulation and visualize the internal activities and thoughts that contain meaningful semantic multimedia concepts; (ii) designing external stimuli with specified semantics for EEG-based recognition and detection, and hence providing more challenging learning environment in order to research new deep learning models.

Acknowledgment

The authors wish to acknowledge the financial support from: (i) Natural Science Foundation China (NSFC) under the Grant No. 61620106008; and (ii) The second round of Shenzhen University Research Foundation funding (2018-2020).

References

[1] S. Moon and J. Lee, "Implicit analysis of perceptual multimedia experience based on physiological response: A review," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 340–353, Feb 2017.

[2] J. Kulasingham, V. Vibujithan, and A. De Silva, "Deep belief networks and stacked autoencoders for the p300 guilty knowledge test," in *Biomedical Engineering and Sciences (IECBES), 2016 IEEE EMBS Conference on*. IEEE, 2016, pp. 127–132.

[3] F. Cong, V. Alluri, A. K. Nandi, P. Toiviainen, R. Fa, B. Abu-Jamous, L. Gong, B. G. W. Craenen, H. Poikonen, M. Huotilainen, and T. Ristaniemi, "Linking brain responses to naturalistic music through analysis of ongoing eeg and stimulus features," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1060–1069, Aug 2013.

[4] X. Liu, X. Tao, M. Xu, Y. Zhan, and J. Lu, "An eeg-based study on perception of video distortion under various content motion conditions," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[5] G. R. Muller-Putz and G. Pfurtscheller, "Control of an electrical prosthesis with an ssvep-based bci," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 361–364, 2008.

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

[7] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, R. Boots, and B. Benatallah, "Cascade and parallel convolutional recurrent neural networks on eeg-based intention recognition for brain computer interface," in *AAAI Conference on Artificial Intelligence*, 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16107

[8] 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.

[9] J. Wang, E. Pohlmeyer, B. Hanna, Y.-G. Jiang, P. Sajda, and S.-F. Chang, "Brain state decoding for rapid image retrieval," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 945–954.

[10] 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 *International Conference on Multimedia Modeling*. Springer, 2015, pp. 447–457.

[11] 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.

[12] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2image: Converting brain signals into images," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1809–1817.

[13] Y. Zhong and Z. Jianhua, "Cross-subject classification of mental fatigue by neurophysiological signals and ensemble deep belief networks," in *Control Conference (CCC), 2017 36th Chinese*. IEEE, 2017, pp. 10 966–10 971.

[14] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted boltzmann machines," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 25, no. 6, pp. 566–576, 2017.

[15] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep feature learning for eeg recordings," *arXiv preprint arXiv:1511.04306*, 2015.

[16] T. Ogawa, Y. Sasaka, K. Maeda, and M. Haseyama, "Favorite video classification based on multimodal bidirectional lstm," *IEEE Access*, vol. 6, pp. 61 401–61 409, 2018.

[17] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4503–4511.

[18] A. Gogna, A. Majumdar, and R. Ward, "Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2196–2205, 2017.

[19] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[20] P. Tirupattur, Y. S. Rawat, C. Spampinato, and M. Shah, "Thoughtviz: Visualizing human thoughts using generative adversarial network," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: ACM, 2018, pp. 950–958. [Online]. Available: http://doi.acm.org/10.1145/3240508.3240641

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969125

[22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6114

[23] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah, "Generative adversarial networks conditioned by brain signals," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 3430–3438.

[24] S.-h. Zhong, A. Fares, and J. Jiang, "An attentional-lstm for improved classification of brain activities evoked by images," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019, pp. 1295–1303. [Online]. Available: http://doi.acm.org/10.1145/3343031.3350886

[25] N. Al-Hadithi, A. Al-Imam, M. Irfan, M. Khalaf, and S. Al-Khafaji, "The relation between cerebral dominance and visual analytic skills in iraqi medical students, a cross sectional analysis," *Asian Journal of Medical Sciences*, vol. 7, no. 6, pp. 47–52, Oct. 2016. [Online]. Available: https://www.nepjol.info/index.php/AJMS/article/view/15205

[26] D. Li, T. Yao, L. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 416–428, Feb 2019.

[27] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "High-quality image captioning with fine-grained and semantic-guided visual attention," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1681–1693, July 2019.

[28] P. Rodriguez Lopez, D. Velazquez Dorta, G. Cucurull Preixens, J. M. Gonfaus Sitjes, F. X. Roca Marva, and J. Gonzalez, "Pay attention to the activations: a modular attention mechanism for fine-grained image recognition," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[29] H. Wu, X. Ma, and Y. Li, "Convolutional networks with channel and stips attention model for action recognition in videos," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[30] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1412–1424, June 2019.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. [Online]. Available: http://arxiv.org/abs/1511.06434

[33] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248–255.

[34] N. Sheehy, *Electroencephalography: Basic Principles, Clinical Applications and Related Fields.* Urban and Schwarzenberg, 1982.

[35] N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience," *Frontiers in systems neuroscience*, vol. 2, p. 4, 2008.

[36] N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini, "Matching categorical object representations in inferior temporal cortex of man and monkey," *Neuron*, vol. 60, no. 6, pp. 1126–1141, 2008.

[37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6626–6637. [Online]. Available: http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium.pdf

[38] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Thirty-sixth International Conference on Machine Learning (ICML)*, 2019.

[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. USA: Curran Associates Inc., 2016, pp. 2234–2242. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157096.3157346

[40] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah, "Decoding brain representations by multimodal learning of neural activity and visual features," *CoRR*, vol. abs/1810.10974, 2018. [Online]. Available: http://arxiv.org/abs/1810.10974

[41] J. Jiang, A. Fares, and S. Zhong, "A context-supported deep learning framework for multimodal brain imaging classification," *IEEE Transactions on Human-Machine Systems*, pp. 1–12, 2019.

[42] P. Mukherjee, A. Das, A. K. Bhunia, and P. P. Roy, "Cogni-net: Cognitive feature learning through deep visual perception," *CoRR*, vol. abs/1811.00201, 2018. [Online]. Available: http://arxiv.org/abs/1811.00201

[43] R. Li, J. S. Johansen, H. Ahmed, T. V. Ilyevsky, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, "Training on the test set? an analysis of spampinato et al. [31]," *CoRR*, vol. abs/1812.07697, 2018. [Online]. Available: http://arxiv.org/abs/1812.07697

[44] J. Wang, V. L. Cherkassky, Y. Yang, K. min Kevin Chang, R. Vargas, N. Diana, and M. A. Just, "Identifying thematic roles from neural representations measured by functional magnetic resonance imaging," *Cognitive Neuropsychology*, vol. 33, no. 3-

4, pp. 257–264, 2016, pMID: 27314175. [Online]. Available: https://doi.org/10.1080/02643294.2016.1182480

[45] S. Makeig, M. Westerfield, T. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. Sejnowski, "Electroencephalographic sources of visual evoked responses," *Science*, vol. 295, pp. 690–694, 2002.

Jianmin Jiang received PhD from the University of Nottingham, UK, in 1994. From 1997 to 2001, he worked as a full professor of Computing at the University of Glamorgan, Wales, UK. In 2002, he joined the University of Bradford, UK, as a Chair Professor of Digital Media, and Director of Digital Media & Systems Research Institute. He worked at the University of Surrey, UK, as a full professor during 2010-2014 and a distinguished chair professor (1000-plan) at Tianjin University, China, during 2010-2013. He is currently a Distinguished Chair Professor and director of the Research Institute for Future Media Computing at the College of Computer Science & Software Engineering, Shenzhen University, China. He was a chartered engineer, fellow of IEE, fellow of RSA, member of EPSRC College in the UK, and EU FP-6/7 evaluator. His research interests include, image/video processing in compressed domain, digital video coding, medical imaging, computer graphics, machine learning and AI applications in digital media processing, retrieval and analysis. He has published around 400 refereed research papers.

Ahmed Fares received his Ph.D. degree from Department of Computer Science and Engineering, Egypt-Japan University of Science and Technology (E-JUST), in 2015. Currently, he is a Postdoctoral Researcher in Research Institute for Future Media Computing at the College of Computer Science & Software Engineering, Shenzhen University, and an Assistant Professor in Department of Electrical Engineering, Computer Engineering branch at the Faculty of Engineering at Shoubra, Benha University. His research interests include, brain science, cognitive science, computational modeling, theoretical computer science, and machine learning

Sheng-hua Zhong received her Ph.D. from Department of Computing, The Hong Kong Polytechnic University in 2013. She worked as a Postdoctoral Research Associate in Department of Psychological & Brain Sciences at The Johns Hopkins University from 2013 to 2014. Currently, she is an Assistant Professor in College of Computer Science & Software Engineering at Shenzhen University in Shenzhen. Her research interests include multimedia content analysis, cognitive science, psychological and brain science, and machine learning.