Learning Spatiotemporal Graph Representations for Visual Perception Using EEG Signals

Jenifer Kalafatovich, Minji Lee, and Seong-Whan Lee, *Fellow, IEEE*

***Abstract*— Perceiving and recognizing objects enable interaction with the external environment. Recently, decoding brain signals based on brain-computer interface (BCI) that recognize the user's intentions by just looking at objects has attracted attention as a next-generation intuitive interface. However, classifying signals from different objects is very challenging, and in practice, decoding performance for visual perception is not yet high enough to be used in real environments. In this study, we aimed to classify single-trial electroencephalography signals evoked by visual stimuli into their corresponding semantic category. We proposed a two-stream convolutional neural network to increase classification performance. The model consists of a spatial stream and a temporal stream that use graph convolutional neural network and channel-wise convolutional neural network respectively. Two public datasets were used to evaluate the proposed model; (i) SU DB (a set of 72 photographs of objects belonging to 6 semantic categories) and MPI DB (8 exemplars belonging to two categories). Our results outperform state-of-the-art methods, with accuracies of 54.28 ± 7.89% for SU DB (6-class) and 84.40 ± 8.03% for MPI DB (2-class). These results could facilitate the application of intuitive BCI systems based on visual perception.**

***Index Terms*— Visual perception, electroencephalography (EEG), convolutional neural network (CNN), brain–computer interface (BCI).**

Manuscript received 13 January 2022; revised 12 June 2022 and 16 September 2022; accepted 8 October 2022. Date of publication 26 October 2022; date of current version 30 January 2023. This work was supported in part by Institute for Information & Communications Technology Promotion (IITP) grants funded by the Korea government under grants 2015-0-00185 (Development of Intelligent Pattern Recognition Softwares for Ambulatory Brain-Computer Interface), 2017-0-00451 (Development of BCI based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning), 2019-0-00079 (Artificial Intelligence Graduate School Program, Korea University) and 2021-0-02068 (Artificial Intelligence Innovation Hub). *(Corresponding author: Seong-Whan Lee.)*

Jenifer Kalafatovich and Seong-Whan Lee are with the Department of Artificial Intelligence, Korea University, Seoul 02841, South Korea (e-mail: jenifer@korea.ac.kr; sw.lee@korea.ac.kr).

Minji Lee is with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea (e-mail: minjilee@korea.ac.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2022.3217344, provided by the authors.

Data is available on-line at https://github.com/JeniferK28/Learning-Spatiotemporal-Graph-Representations-for-Visual-Perception-using-EEG-Signals

Digital Object Identifier 10.1109/TNSRE.2022.3217344

I. Introduction

PERCEPTION and recognition of objects are essential for the interaction with the external environment and with other people [1]. They facilitate the planning and execution of motor action, which requires information regarding the location and material properties of objects [2]. It has been proved that the identification objects can be accessed rapidly when presented visually compared to other modalities such as text [3]. In this line, many studies concluded that the occipital cortex is the one responsible for visual information processing [4]. Information related to object identification and the corresponding semantic category is reflected by changes in brain activity, which can be extracted in less than 200 ms depending on the semantic category of the presented stimulus [5]. Moreover, decoding conceptual information has gained interest since it can be applied to a brain-computer interface (BCI) system that aims to transform lexical concepts into written or spoken output [6].

Humans can recognize objects in a matter of milliseconds [7]. Brain signals have been widely studied in order to understand the neural mechanisms involved in this ability [8]. As a result, changes in brain activity have been associated with the presentation of a stimulus of a certain semantic category [9]. Haxby et al. [10] recorded functional magnetic resonance imaging (fMRI) signals when subjects were presented with stimuli of different semantic categories (human faces, cats, house chairs, scissors, shoes, bottles, and nonsense images). The similarity between patterns was analyzed using correlations of brain responses. It was reported that different brain regions of the ventral temporal cortex are preferentially activated, and a distinct pattern of brain responses is elicited depending on the semantic category of the presented images [10], [11]. However, the differences in brain signals corresponding to visual stimuli from different semantic categories are not clear.

Many studies of visual perception in humans have been actively performed using electroencephalography (EEG) signals due to their suitability for the development of BCI [12], [13]. Some of them focused on event-related potentials (ERP) when comparing evoked signals in response to images of different semantic categories, especially for faces versus objects [14]. A previous study found a significant negative activity (approximately 120–200 ms) after stimulus onset, depending on the stimulus category [15]. Philiastides et al. [16]

presented subjects with images corresponding to faces and cars. The authors identified two maximally discriminating facial components. An early component was found at 170 ms after stimulus onset and a late component was found at least 130 ms after the first one. Simanova et al. [6] analyzed discriminant information for two object categories (animal and tool class). A positive and a negative ERP component were elicited at 110 ms and 160 ms after stimulus onset respectively; these were largest over the inferotemporal and occipital electrodes. Additionally, the deflection over the frontal electrodes was less negative for the animal than for tool trials.

Brain activity has been modeled using graph theory [17]. Moreover, the interest in graph neural networks in magnetoencephalography (MEG), fMRI, and EEG studies had been increasing, yet it still represents a challenge [18]. Previous works had applied graph convolutional neural networks to EEG signals [19]. Some of these explored electrode distance [20], [21], while others used functional connectivity values [22] such as phase-locking value (PLV) for the construction of graphs. They proposed that functional connectivity represents the interactions between brain regions that occur in the brain [23], whereas neighbor electrodes connections represent the interactions inside a specific brain region. The above mention studies focused on feature extraction while modeling connections between electrodes but failed to extract channel-wise features.

In this study, we attempted to decode the semantic category of the presented stimuli using single-trial EEG signals. First, we analyzed time-domain features of the brain signals and the functional connectivity between trials of different semantic categories. We found significant differences in both time-domain and functional connectivity for different classes, as a result, we concluded that these features could be used for classification. Second, we proposed a two-stream convolutional neural network (TSCNN) for classifying visual perception. Graph convolutional neural network (GCNN) is used to extract spatial relation between electrodes and a convolutional network (CNN) is used to extract channel-wise features; the output of both networks is concatenated and classification is performed. EEG signals are used to construct a graph that is later input to the GCNN. Unlike other studies, we combine local and distant connections of the brain by constructing the adjacency matrix using electrode distance and functional connectivity. There are multiple ways to estimate functional connectivity, a recent study proves that weighted phase lag index (wPLI) attenuated the influence of noise contamination [24] and volume conduction [25], therefore, we decided to use wPLI for estimating the connectivity between a pair electrodes. The CNN extracts channel-wise features, specifically time-domain features that could be ignored by the GCNN since it gives more importance to the spatial relation between electrodes. Our findings lead to better classification accuracy than those in previous methods and demonstrate that it is possible to classify single-trial EEG signals generated during the representation of visual stimuli into different semantic categories with significantly higher accuracies.

The contributions of this study can be summarized as follows.

- We analyzed time-domain features of the brain signals and the functional connectivity between trials of different semantic categories. Our findings support their use for classifying EEG signals into different semantic categories.
- We proposed a framework that combines time-domain features with spatial relations between electrodes by simulating connections of the brain. Different from other methods, we simulate local and distant connections simultaneously by using electrodes distance and functional connectivity to learn graph representations.
- We demonstrated the effectiveness of our method by improving classification accuracies when compared to networks that use time-domain features or spatial relations between electrodes separately. Moreover, our model outperformed the existing state-of-the-art methods for decoding human visual perception and classifying EEG signals.

II. Related Works

Multiple studies have attempted to classify observed objects from brain signals using fMRI [26], [27], MEG [28], [29], and EEG [1], [30], [31], [32], [33]. fMRI has high spatial resolution when acquiring brain signals [34], therefore decoding performance tends to be higher than when using EEG. However, due to its low time resolution, its use in BCI systems is limited. In contrast, EEG has been applied to a variety of BCI systems due to its high time resolution and portability [35].

Initially, traditional machine learning techniques were used to classify brain signals elicited during object perception. Above all, classifying brain signals simply by looking at objects represent a challenge and it is more complex than using clear visual stimuli, such as steady-state visually evoked potential using differences in frequencies. Wang et al. [9] used EEG signals to investigate brain activity patterns related to the encoding of semantic category information. Subjects were presented with images of four semantic categories. ERP was used as a feature extraction method and classification was performed using fisher linear discriminant analysis (LDA). An accuracy of 67.87% was reported for 4-class classification. Kaneshiro et al. [1] aimed to classify visual stimuli in semantic categories (6-class) and individual exemplars (72-class). Principal component analysis was applied to raw EEG signals to reduce dimensionality and LDA was used for classification. The reported classification accuracies for semantic categories and individual exemplars were 40.68% and 14.46%, respectively.

Deep learning algorithms have been successfully applied in many areas such as the classification of images, text, and even biosignals [36]. However, the application of deep learning algorithms to brain signals has not been able to achieve the best performance yet [37]. Traditional machine learning algorithms use a feature extractor (dimensional reduction) and a classifier, therefore it is important to consider multiple design options. In contrast, most deep learning algorithms adopt an

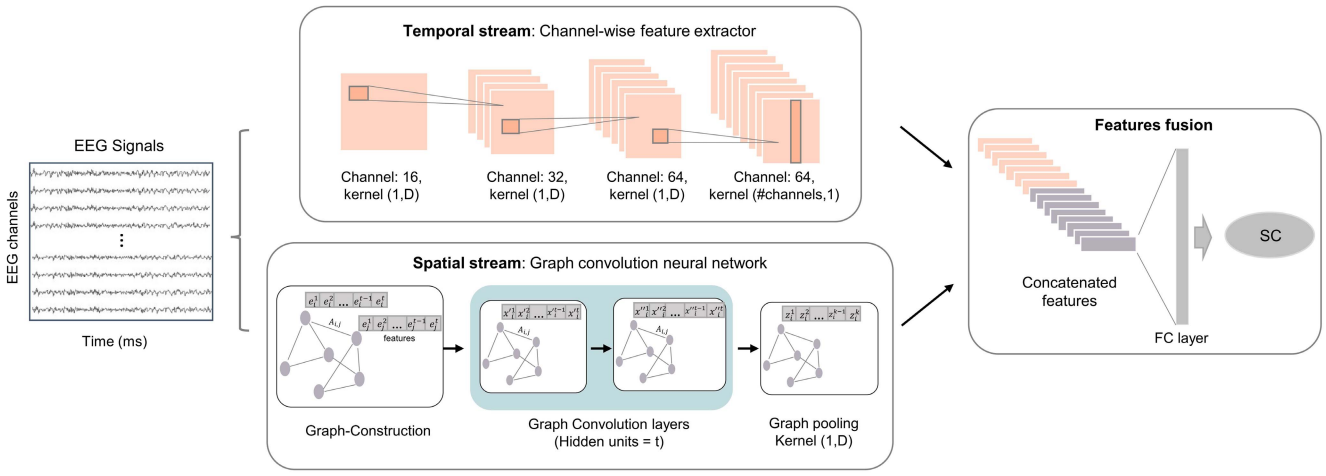

Fig. 1. Proposed framework for the two-stream convolutional network (TSCNN). EEG signals (channels × time points) are input to the network. TSCNN uses three modules; a channel-wise feature extraction (one-stream convolutional neural network - OSCNN), a graph convolutional neural network (GCNN), and a features fusion sub-network. The time-domain signal is used to extract channel-wise time features and to construct a graph that later is input to the GCNN. The sub-network classifies semantic categories by concatenating outputs from the OSCNN and GCNN and inputting them to a fully connected layer. (FC layer: Fully-connected layer, $D = 5$ or 20 and #channels = 124 or 60 - when SU DB or MPI DB is used respectively, t: time points, x': output of the first graph convolution layer, x'' : output of the second graph convolution layer, z: output of the pooling layer, $k : t/D$, SC: semantic category).

end-to-end approach where the pre-processed signal is used as input, and the model performs the feature extraction and classification.

Zheng et al. [30] used EEG signals during the presentation of different images; 40 images were presented 50 times each. The authors applied the Swish activation function to a long short-term memory network (LSTM) over the pre-processed EEG signals; as a result, an accuracy of 97.13% was obtained. Another work used an attention-based bidirectional LSTM network and applied it to the above-mentioned dataset [31]. They incorporate two attention strategies into the traditional LSTM module. The attention gate replaces the forget gate on traditional LSTM and reduces the training parameter. While the attention weighting method is applied to the output of the LSTM module. An accuracy of 99.50% was reported for 40-class. However, in the aforementioned works, stimuli were presented in blocks and in an unrandomized order; in other words, trials of the same stimuli were presented consecutively. Although this justifies the high classification accuracies obtained, these methods cannot be applied to a real-world EEG system. To this end, Ahmed et al. [32] studied the effects of randomizing the presentation of the stimulus on an object perception task. They presented 40 images (same stimuli as in [30] and [31]) to one subject, and classified the evoked EEG signals. An accuracy of 5.4% was obtained using a recurrent neural network over the pre-processed EEG signals. This represents a huge difference with [30] and [31], and showed the importance of randomizing the stimuli during data recording. A more recent work uses multi-headed self-attention and temporal convolution to classify visual stimulus [38]; the transformer capture inter-region interactions while the convolution filters learn temporal patterns. They reported an accuracy of $52.33 \pm 8.28\%$ for 6-class.

As mentioned before, GCNN has been used in EEG studies. Zhang et al. [21] proposed a GCNN based on functional connectivity; specifically, PLV connectivity values applied to emotion recognition. They obtained 84.35% as classification accuracy for three classes. Similarly, Jin et al. [39] used PLV values and graph representations to classify motor imagery tasks. Another work used the Pearson correlation coefficient for the graph construction [40]. They classified motor imagery tasks (4-class) and obtained 93.05% and 96.24% for two different datasets. Song et al. [20] used a graph neural network and explored electrode distance for emotion recognition. They obtained 90.4% and 79.95% as classification accuracy for subject-dependent and subject-independent classification. Since GCNN has proved its efficacy in EEG studies, we utilized GCNN along with CNN to classify visual perception into their semantic categories. Moreover, we model local and distant connections that take place in the brain by using wPLI and electrode distance to construct the graphs.

III. Proposed Method

We proposed the TSCNN to classify the presented visual stimuli into different semantic categories. Fig. 1 depicts the proposed framework; in particular, we used CNN to extract channel-wise time features and GCNN to model the relation between electrodes. The model receives as input the pre-processed EEG data (channels×time points - SU DB: 124×32; MPI DB: 60×250). The temporal stream extracts features from the input directly, meanwhile, the spatial stream constructs graphs using the pre-processed signals and the adjacency matrix (electrodes' distance and functional connectivity values).

A. Temporal Stream: Channel-Wise Feature Representation

The temporal stream (one-stream CNN - OSCNN) consists of four convolutional layers with an exponential linear unit (ELU) as the activation function and a fully connected layer with a softmax activation for the classification. Additionally,

TABLE I
TEMPORAL STREAM (OSCNN) DESCRIPTION

Layer	Operation	Kernel Size	Feature Map	Strides
1	Convolution Batch Normalization	(1, 5) (1, 20)	16	(1, 1) (1, 1)
2	Convolution Batch Normalization	(1, 5) (1, 20)	32	(1, 1) (1, 1)
	Max-pooling	(1, 2) (1, 5)		(1, 2) (1, 5)
3	Dropout Convolution Batch Normalization	(1, 5) (1, 20)	64	(1, 1) (1, 1)
	Max-pooling	(1, 2) (1, 5)		(1, 2) (1, 5)
4	Dropout Convolution Batch Normalization	(124, 1) (60, 1)	64	(1, 1) (1, 1)

dropout ($p = 0.25$) and batch normalization were used to avoid overfitting of the model [37], [41]. Detailed network description can be observed in Table I; as dimensions of both datasets differ from each other (sampling rate, number of electrodes, and number of classes), parameters such as kernel size and stride are set different for each dataset (In Table I, the first line indicates kernel size and stride for SU DB and the second line indicates kernel size and stride for MPI DB). The network's parameters were selected from a pool of values using grid search. Different kernel sizes were tested (For SU DB: (1, 2), (1, 4), (1, 5), (1, 10); for MPI DB: (1, 5), (1, 10), (1, 20), (1, 40)); and the kernel size that produced the best results was selected. For the pooling method, max-pooling and mean pooling were evaluated; however, we decided to use max-pooling as the pooling method since better results were obtained.

B. Spatial Stream: Relation Between Electrodes

We used the time domain EEG signals to construct graphs, a graph is defined as $G = (V, E, A)$ where V represents a set of nodes, E denotes a set of edges of the graph, and A $\in \mathbb{R}^{n_e \times n_e}$ is the adjacency matrix that represents the relation between any pair of nodes (n_e: number of electrodes). In this study, we consider each electrodes as a node, $V = \{e_i\}$, $i \in [1, n_e]$, and time points as nodes features, $e_i = \{e_i^1, e_i^2, e_i^3, \ldots, e_i^{t-1}, e_i^t\}$.

The adjacency matrix is constructed using electrodes distance and functional connectivity following Equation 1, where $\vee$ represents the OR operator; A_L and A_{fc} are defined in Equations 2 and 4, and their shape is $n_e \times n_e$.

$$A = A_L \vee A_{fc} \tag{1}$$

Distance between electrodes is calculated using Euclidean distance $L_{i,j} = \sqrt{(e_{xi} - e_{xj})^2 + (e_{yi} - e_{yj})^2 + (e_{zi} - e_{zj})^2}$, where (e_{xi}, e_{yi}, e_{zi}) represents the locations of channel e_i (since electrodes were distributed on a cap an following the manufacturer layout, channel locations were known); next A_L is calculated following Equation 2, where τ_1 represents a threshold value, which is used to identify neighbor electrodes ($\tau_1 = 0.2$).

$$A_{L_{i,j}} = \begin{cases} 1, & if\ L_{i,j} \leq \tau_1\ \ \&\ \ i \neq j, \\ 0, & otherwise \end{cases} \tag{2}$$

For the functional connectivity, we used w_{PLI} coefficient which is calculated by applying the Fourier transform with a Hanning window for each epoch and each subject separately [23] (see Equation 3). w_{PLI} measure the extent to which phase angle differences between two time series x(t) and y(t) are distributed towards positive or negative parts of the imaginary axis in the complex plane [42].

$$w_{PLI} = \left|\frac{\sum_{t=1}^{n} |imag(S_{xy,t})| sgn(imag((S_{xy,t}))}{\sum_{t=1}^{n} |imag(S_{xy,t})|}\right| \tag{3}$$

where $imag$ is the imaginary component of the cross-spectrum $S_{xy,t}$ of two signals $x(t)$ and $y(t)$ at trial t, and sgn is the sigmun function [42]. Finally A_{fc} matrix is calculated using w_{PLI} values following Equation 4; where τ_2 represents a threshold value, which is used to consider only strong connections between electrodes ($\tau_2 = 0.8$).

$$A_{fc_{i,j}} = \begin{cases} 1, & if\ w_{PLI(i,j)} \geq \tau_2\ \ \&\ \ i \neq j, \\ 0, & otherwise \end{cases} \tag{4}$$

The spatial stream (GCNN) is composed of two graph convolutional layers (hidden features = 32 or 500) and graph average pooling with a kernel of (1, 5) and (1, 20) for SU DB and MPI DB respectively; as an alternative to the average pooling layer; max-pooling layer was evaluated, however, better results were obtained using average pooling layer. Similar to channel-wise feature extraction different kernel sizes were tested and grid search was used to select optimal values. The convolution operation on the GCNN is performed following Equation 5.

$$x_i' = G(W_1 x_i + W_2 \sum A_{i,j} \cdot x_j) \tag{5}$$

where W_1 and W_2 are the weights that regulate the influence of all electrodes and are learned by the network. $G()$ indicates the activation function, after performing grid search (between ELU, ReLU, and sigmoid function); ELU was selected. x_i represents the node to the be analyze while x_j represents the other nodes (electrodes); i and $j \in [0, n_e]$. The shape of x_i and x_j is $1 \times t$ (t: number of time points). The training process is presented in Algorithm 1.

C. Feature Fusion for Classification

The output of the temporal and spatial stream are concatenated and input to the sub-network, which classifies the semantic category of the shown stimuli. High-level features in the time domain and the relation between electrodes obtained through the GCNN are used. When training the model, both streams were trained simultaneously. The model is trained using Adam optimizer [43] and minimizing the cross-entropy loss function, the learning rate is set to 0.005.

Algorithm 1 Training Process.

Input: EEG data set (X, Y)

- $X = \{x_d\}_{k=1}^{D}, x \in \mathbb{R}^{E \times T}$: set of single trial EEG signal, where D is the number of trials with E number of electrode channels and T time length
- $Y = \{y_{sc}\}_{k=1}^{D}$: class label, where y_{sc} is the label of semantic class

Output: Trained TSCNN model

Divide EEG data into k folds
for z = 1 to k **do**
  Set fold k_z as test set (X_{test}) and remaining folds as training set (X_{train})
  for d = 1 to n_t (n_t: Number of trials in training set) **do**
    Calculate electrode distance matrix $L_{i,j}^{d}$, where i and j $\in \{1, E\}$
    if $L_{i,j}^{d} \leq \tau_1 \ \& \ i \neq j$ **then**
      $A_{L_{i,j}}^{d} = 1$
    else $A_{L_{i,j}}^{d} = 0$
    end if
    Calculate functional connectivity for each trial $wPLI_{i,j}^{d}$
    if $wPLI_{i,j}^{d} \geq \tau_2 \ \& \ i \neq j$ **then**
      $A_{fc_{i,j}}^{d} = 1$
    else $A_{fc_{i,j}}^{d} = 0$
    end if
    Calculate adjacency matrix for each trial $A^d = A_L^d \vee A_{fc}^d$ and construct graph $G^d(x^d, A^d)$
  end for
  for iterations = 1 to n_i (n_i: Number of iterations) **do**
    $X_1' = GCNN(G_{train}, A)$ /* Graph convolutional neural network */
    $X_2' = OCNN(X_{train})$ /* Convolutional neural network */
    $Y' = FC(X_1' \parallel X_2')$ /* Fully connected layer */
    Calculate loss term
    Minimize loss values
  end for
end for
Grid search of optimal parameters

IV. Experimental Results

A. Datasets and Evaluation Metrics

We used two datasets that recorded EEG signals during visual perception tasks; Stanford University dataset (SU DB) [1] and Max-Planck Institute dataset (MPI DB) [6]. The comparison between the two datasets is shown in Table II.

1) SU DB: The first dataset was published by Kaneshiro et al. [1]. EEG signals from ten healthy subjects (3 females, 21–57 years, 1 left-handed subject) with normal color vision and normal or corrected-to-normal vision were measured. The experimental paradigm comprised the presentation of photographs from one of the following six semantic categories: human body (HB), human face (HF), animal body (AB), animal face (AF), fruits or vegetables (FV), and inanimated objects (IO). In total, 72 images set against a mid-gray background were presented, with 12 different images per semantic category (Supplementary Fig. 1). An image was presented for 500 ms followed by a blank gray screen for 750 ms. Each image was shown 72 times as follows: the experiment was divided into two sessions and each session consisted of three blocks, each of which included 864 trials (each image was presented 12 times randomly) with short breaks after every 36 trials. EEG signals were filtered between 1 and 25 Hz using a high-pass fourth-order Butterworth filter and a low-pass eight-order Chebyshev Type I filter, respectively; finally, signals were temporally downsampled to 62.5 Hz.

TABLE II
Datasets Description

	SU DB [1]	MPI DB [6]
# of channels	124	60
# of subjects	10	20
Sampling rate	62.5 Hz	500 Hz
Stimuli presentation	Visual stimuli (color photographs)	Visual stimuli (no color drawings)
Class	6 semantic categories	2 semantic categories
# of trials	864 per category	320 per category
Cue length	500 ms	300 ms

2) MPI DB: The second dataset was published by Simanova et al. [6]. EEG signals from 24 subjects (14 females, 18-28 years, all right-handed, with no neurological disorders), from which four subjects were selected as a pilot group, were recorded. The presented stimulus belongs to three semantic categories, relevant categories (animals (A) and tool (T)), and task category (clothing or vegetable), each relevant category contained four exemplars (animals classes: cow (C), bear (B), lion (L), ape (A); tools classes: ax (Ax), scissors (Sc), comb (Co), pen (Pe)) (Supplementary Fig. 2). All exemplars were shown in three modalities: auditory, visual, and orthographical (in this study, we used signals from visual modality and relevant categories). Each of the stimuli belonging to a relevant category was presented 80 times. The stimulus was presented at 300 ms and followed by a black screen during 1,000–1,200 ms. Stimuli were randomized and presented in 12 blocks with a short break between blocks. Finally, signals were filtered below 1 Hz and above 30 Hz with a sampling rate of 500 Hz.

3) Performance Metrics: Classification was performed for each subject independently (subject-dependent approach). All models were trained and evaluated using a 10-fold cross-validation method within the subject. Data from one subject was randomly partitioned into ten subsets called folds. Nine folds were used as training data and the remaining fold was used as the test set. This process was performed until all the folds were used as the test set [44]. We used accuracy, precision, sensitivity, specificity, and F1-scores as performance metrics.

B. Changes in EEG Signals According to Visual Stimuli

1) Event-Related Potentials: Previous studies have shown differences in the elicited brain activity between semantic categories depending on the brain region analyzed [14], [15].

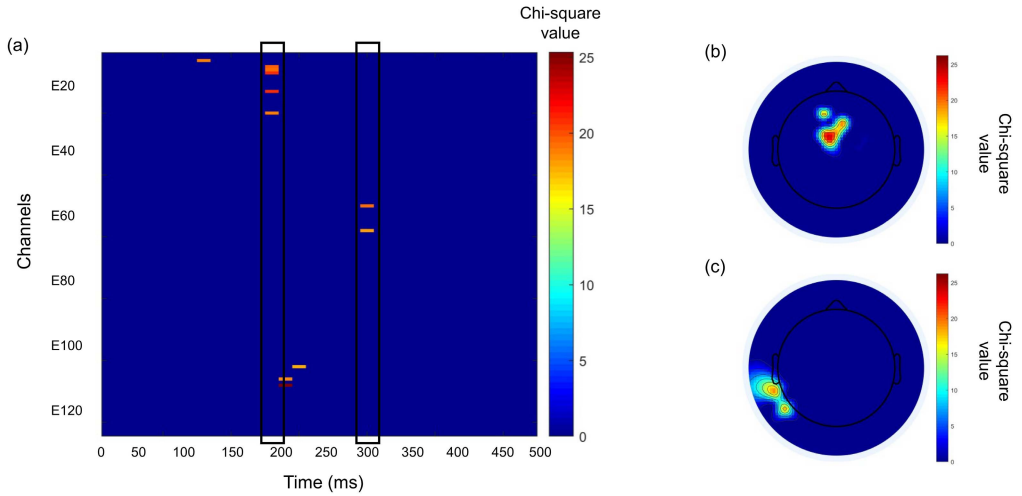

Fig. 2. Differences in event-related potentials between semantic categories of SU DB (6-class). Chi-square values across all channels are shown in (a). Scalp plots for the time intervals that showed significant differences are illustrated, specifically at (b) 190 ms and (c) 300 ms. If p-value is greater than 0.05, we set chi-square values to zero to clarify the interval with significant differences ($p < 0.05$ with Bonferroni correction).

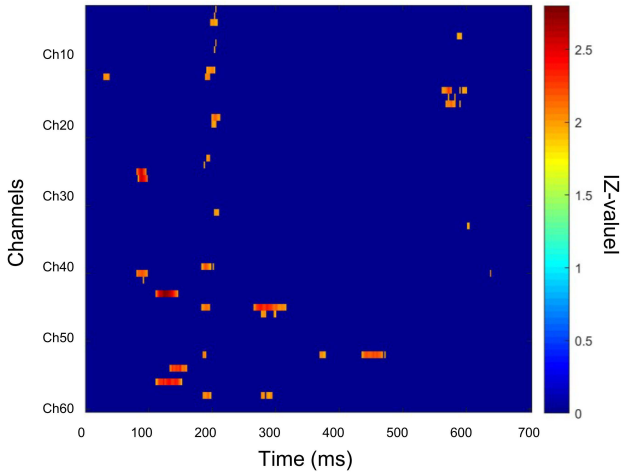

Fig. 3. Differences in event-related potentials for MPI DB (2-class). |Z-values| are shown across all channels. If p-value is greater than 0.05, we set z-scores to zero to clarify the interval with significant differences ($p < 0.05$ with Bonferroni correction).

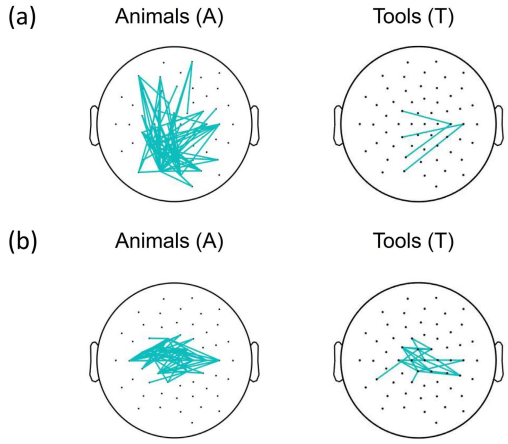

Fig. 4. Average functional connectivity (wPLI) values across trials of the same semantic category for two representative subjects of MPI DB. (a) Sub18 (b) Sub15. (If $wPLI \geq 0.8$, we draw a line between the evaluated electrodes.)

Therefore, EEG signals were divided into different semantic categories and averaged across trials within the stimulus class. Statistical analyses were performed to determine significant differences. The normality of data at all the time points was checked using the Shapiro-Wilk test, the null hypothesis was rejected, therefore non-parametric tests were used for all the comparisons. Comparison between semantic categories at each time point was performed to determine significant differences. For SU DB, Kruskal-Wallis test (non-parametric one-way analysis of variance) was conducted on the average amplitude of the EEG signals from each participant. For the comparison of trials corresponding to the semantic category of animal and tool in MPI DB, Wilcoxon rank-sum test was used. The significance level was set at $p = 0.05$ with Bonferroni correction.

Fig. 2 illustrates the statistical results of the comparison between the six semantic categories for all channels corresponding to SU DB. Kruskal-Wallis test revealed significant differences in different brain regions; relevant differences were found at approximately 190 ms and 300 ms. Particularly, a significant difference was observed over frontal and occipito-temporal regions.

For MPI DB, the comparison between the two semantic categories (animal and tool) revealed significant differences at 80–100 ms in temporal and occipital regions; 110–150 ms in occipital; 180–200 ms in temporal regions; 260–320 ms in occipito-temporal regions; and 430–470 ms in temporal regions (Fig. 3).

2) Functional Connectivity: Functional connectivity was calculated following [23] for each epoch and each subject separately. wPLI values were later used to construct graphs along with the distance between electrodes. Fig. 4 depicts wPLI values for two representative subjects of MPI DB. There were differences between functional connectivity for animal class and tool class for both subjects. The wPLI values were also calculated for SU DB (see Fig. 5); results showed differences in the connections between all semantic categories; however,

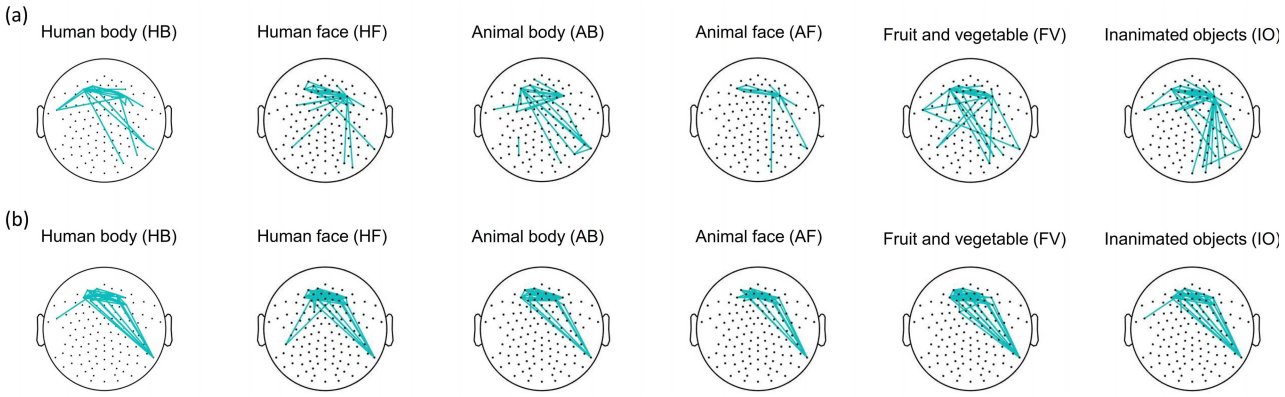

Fig. 5. Average functional connectivity (wPLI) values across trials of the same semantic category for two representative subjects of SU DB. (a) Sub06 (b) Sub08. (If $wPLI \geq 0.8$, we draw a line between the evaluated electrodes.)

TABLE III
CLASSIFICATION ACCURACY USING TSCNN FOR SU DB

Subject	Accuracy (%)	F1-score	Precision	Sensitivity	Specificity
Sub01	52.19 ± 2.25	0.50 ± 0.09	0.53 ± 0.12	0.53 ± 0.12	0.53 ± 0.08
Sub02	45.74 ± 2.00	0.45 ± 0.09	0.45 ± 0.13	0.45 ± 0.13	0.45 ± 0.07
Sub03	58.38 ± 1.14	0.58 ± 0.09	0.57 ± 0.11	0.57 ± 0.11	0.56 ± 0.07
Sub04	46.04 ± 1.40	0.46 ± 0.12	0.46 ± 0.13	0.46 ± 0.12	0.47 ± 0.09
Sub05	62.86 ± 1.98	0.62 ± 0.08	0.62 ± 0.09	0.63 ± 0.09	0.62 ± 0.06
Sub06	64.80 ± 1.99	0.63 ± 0.08	0.64 ± 0.11	0.64 ± 0.10	0.64 ± 0.07
Sub07	60.28 ± 1.47	0.59 ± 0.10	0.60 ± 0.07	0.60 ± 0.07	0.60 ± 0.07
Sub08	44.00 ± 2.58	0.43 ± 0.09	0.44 ± 0.07	0.44 ± 0.11	0.44 ± 0.08
Sub09	46.04 ± 2.85	0.48 ± 0.14	0.48 ± 0.12	0.48 ± 0.12	0.48 ± 0.07
Sub10	62.49 ± 2.28	0.61 ± 0.11	0.62 ± 0.13	0.62 ± 0.12	0.62 ± 0.07
Average ± Std.	**54.28 ± 7.89**	**0.53 ± 0.07**	**0.54 ± 0.07**	**0.54 ± 0.07**	**0.54 ± 0.07**

TABLE IV
CLASSIFICATION ACCURACY USING TSCNN FOR MPI DB

Subject	Accuracy (%)	F1-score	Precision	Sensitivity	Specificity
Sub01	84.92 ± 5.64	0.84 ± 0.002	0.85± 0.003	0.85 ± 0.002	0.84 ± 0.001
Sub02	95.25 ± 3.32	0.93 ± 0.001	0.93 ± 0.002	0.94 ± 0.001	0.93 ± 0.003
Sub03	81.01 ± 5.85	0.84 ± 0.005	0.83 ± 0.004	0.87 ± 0.007	0.81 ± 0.005
Sub04	90.56 ± 4.35	0.90 ± 0.002	0.90 ± 0.001	0.91 ± 0.002	0.90 ± 0.002
Sub05	93.06 ± 2.24	0.93 ± 0.002	0.93 ± 0.003	0.93 ± 0.001	0.93 ± 0.003
Sub06	72.51 ± 3.75	0.74 ± 0.008	0.75 ± 0.006	0.76 ± 0.007	0.74 ± 0.008
Sub07	78.01 ± 5.32	0.77 ± 0.010	0.78 ± 0.008	0.75 ± 0.009	0.81 ± 0.010
Sub08	84.45 ± 4.75	0.85 ± 0.001	0.86 ± 0.002	0.85 ± 0.002	0.86 ± 0.001
Sub09	91.27 ± 4.53	0.92 ± 0.001	0.94 ± 0.001	0.91 ± 0.001	0.94 ± 0.002
Sub10	85.81 ± 4.35	0.88 ± 0.001	0.89 ± 0.002	0.88 ± 0.001	0.89 ± 0.002
Sub11	90.05 ± 4.45	0.89 ± 0.005	0.91 ± 0.006	0.87 ± 0.005	0.92 ± 0.006
Sub12	72.32 ± 7.03	0.74 ± 0.006	0.75 ± 0.005	0.76 ± 0.005	0.73 ± 0.006
Sub13	75.68 ± 6.32	0.78 ± 0.004	0.78 ± 0.004	0.78 ± 0.004	0.79 ± 0.005
Sub14	93.25 ± 3.84	0.93 ± 0.001	0.93 ± 0.002	0.93 ± 0.001	0.93 ± 0.002
Sub15	71.02 ± 4.58	0.71 ± 0.001	0.72 ± 0.001	0.71 ± 0.002	0.72 ± 0.001
Sub16	85.89 ± 5.76	0.85 ± 0.001	0.86 ± 0.001	0.86 ± 0.001	0.86 ± 0.001
Sub17	78.04 ± 5.35	0.77 ± 0.001	0.78 ± 0.001	0.77 ± 0.002	0.78 ± 0.001
Sub18	96.23 ± 2.89	0.96 ± 0.001	0.97 ± 0.001	0.96± 0.001	0.97 ± 0.001
Sub19	92.27 ± 2.34	0.92 ± 0.003	0.95 ± 0.003	0.90 ± 0.004	0.96 ± 0.002
Sub20	76.52 ± 4.98	0.77 ± 0.001	0.77 ± 0.001	0.78 ± 0.002	0.77 ± 0.001
Average ± Std.	**84.40 ± 8.03**	**0.85 ± 0.07**	**0.85 ± 0.07**	**0.85 ± 0.07**	**0.85 ± 0.07**

there was a strong correlation between frontal and occipital electrodes for all cases. Obtained values were averaged across trials of a specific class, and a threshold value ($wPLI \geq 0.8$) was applied for better visualization.

C. Classification Performance

We performed single-trial EEG classification; the model was implemented using Pytorch framework and trained on NVIDIA GeForce RTX. All models were trained in a maximum of 50 epochs; training and validation accuracies for the highest and lowest subject performance depending on the number of iterations is illustrated in Supplementary Fig. 3. Training accuracies increase with the number of iterations, whereas validation accuracy reaches the maximum value before the 50 epochs in all cases, therefore the used of early-stopping is necessary to avoid overfitting. For the statistical analysis of the classification accuracies, we confirmed the normality of each comparative group using the Shapiro-Wilk test, after normality was proved parametric tests were used for the comparison.

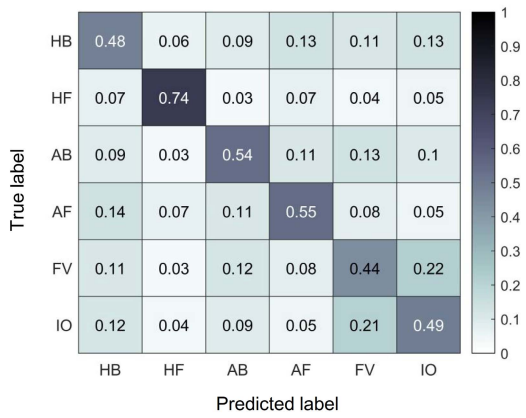

Fig. 6. Average confusion matrix for semantic categories classification of SU DB using TSCNN (HB: human body class, HF: human face class, AB: animal body class, AF: animal face class, FV: fruit or vegetable class, and IO: inanimated objects).

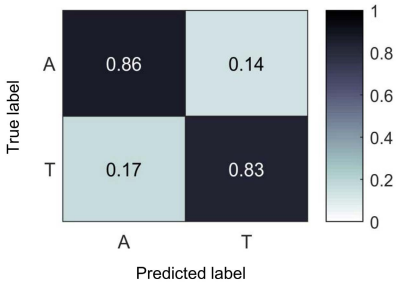

Fig. 7. Average confusion matrix for semantic categories classification of MPI DB using TSCNN (A: animal and T: tool).

1) Classification Using TSCNN: The model was trained and evaluated for each subject independently and results from the ten-folds were averaged. Table III shows average accuracies, F1-scores, precision, sensitivity, and specificity for all subjects in SU DB. The highest and lowest accuracies were obtained for Sub06 and Sub08, respectively. Sub06 obtained accuracies of 64.80 ± 1.99%, and Sub08 obtained accuracies of 44.00 ± 2.58%. The same trend is observed for the other metrics. Table IV shows average results across all ten-folds using TSCNN for all subjects in MPI DB. The highest and lowest accuracies were obtained by Sub18 and Sub15. Sub18 obtained accuracies of 96.23 ± 2.89%, while Sub15 obtained accuracies of 71.02 ± 4.58%.

2) Comparison Between Different Classes: Confusion matrices were averaged across all folds and subjects. Fig. 6 illustrates the averaged confusion matrix for 6-class (SU DB) when using TSCNN. The model mostly confuses FV class with IO class, whereas HF class is the most distinguishable class and FV the less distinguishable class. HB class was mostly confused with AF class, whereas HF class was mostly confused with AF class. Additionally, AB class was mostly confused with FV class. Fig. 7 illustrates the averaged confusion matrix for MPI DB when using TSCNN. Similar accuracies are obtained per class. To evaluate the feature extraction ability of the model, t-sne plots were generated. We extracted the features of the fully-connected layer for an individual fold and used them to graph the plots. As illustrated in Fig. 8, samples of the same class show similarities for both datasets.

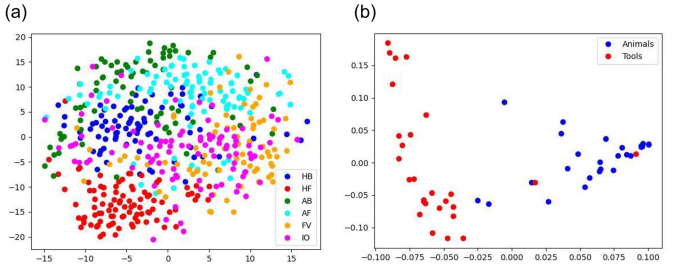

Fig. 8. t-SNE plots for a single fold of a representative subject in each dataset. (a) SU DB (Sub 06) and (b) MPI (Sub 15).

Finally, we performed a binary classification for SU DB; for this, we selected similar classes to MPI DB (AB vs IO). In addition, we classify HF vs IO, since previous works have explored the classification between these two classes. Results are shown in Supplementary Table I and Table II respectively. We observe IO class shows the lowest classification accuracies for both cases. There was a significant differences when classifying AB vs IO (AB: 82.79 ± 7.20%,; IO: 71.72 ± 5.71%, $p < 0.001$); however for HF vs IO case the differences is not significant (HF: 89.11 ± 4.42%,; IO: 86.87 ± 5.05%, $p = 0.14$).

D. Ablation Studies

We performed ablation studies to emphasize the superiority of the proposed method. As mention before, we set $\tau 1$ and $\tau 2$ to 0.2 and 0.8 respectively; and construct the adjacency matrix. To support our decision, we performed classification using GCNN for different values of $\tau 1$ and $\tau 2$; additionally, we evaluate the influence of each of these variables by using either electrodes distance or functional connectivity ($\tau 1$ or $\tau 2$ respectively) on the construction of the adjacency matrix. Results are shown in Supplementary Table III along with the statistical results when compared to the selected values. We observe that significant high performance was obtained when using both $\tau 1$ and $\tau 2$. Even though the performance was slightly higher when using only $\tau 1$ than when using only $\tau 2$ for training the model, there was no significant difference between them.

GCNN design parameters could influence the performance of the model, therefore, we analyze some design options such as the number of layers and the pooling type, and compared the results with the ones on the proposed model. The highest results were obtained when using average pooling and 2 layers (see Supplementary Table IV).

Both branches used on the two-stream CNN are analyzed separately (OSCNN and GCNN). Cross-entropy was used to minimize the loss function and Adam optimizer was used to train both models. Even though different parameters such as kernel size and the number of layers were used, higher accuracies were obtained when using the same parameters as the proposed method. Accuracies corresponding to each class when using the proposed networks (GCNN, OSCNN, and TSCNN) were compared using paired t-test. Fig. 9 depicts the averaged classification accuracies from GCNN, OSCNN, and TSCNN for SU DB and MPI DB across subjects, and the results of the statistical comparison. TSCNN obtained

TABLE V
COMPARISON OF CLASSIFICATION PERFORMANCE BETWEEN PROPOSED AND SOTA

Dataset	Method	Accuracy (%)	p-value*	Precision	p-value*	Sensitivity	p-value*	Specificity	p-value*
SU DB [1]	LDA [1]	40.52 ± 5.62	<0.001	0.41 ± 0.05	<0.001	0.39 ± 0.05	<0.001	0.40 ± 0.05	<0.001
	ShallowConvNet [37]	46.51 ± 6.76	<0.001	0.45 ± 0.09	<0.001	0.45 ± 0.09	<0.001	0.45 ± 0.08	<0.001
	EEGNet [45]	43.83 ± 5.15	<0.001	0.43 ± 0.05	<0.001	0.42 ± 0.05	<0.001	0.43 ± 0.05	<0.001
	LSTM [30]	38.06 ± 1.88	<0.001	0.37 ± 0.04	<0.001	0.38 ± 0.05	<0.001	0.37 ± 0.04	<0.001
	EEG-ConvTransformer [38]	52.33 ± 8.28	-	-	-	-	-	-	-
	TSCNN *(Ours)*	54.28 ± 7.89	-	0.54 ± 0.07	-	0.54 ± 0.07	-	0.54 ± 0.07	-
MPI DB [6]	LDA [1]	76.11 ± 6.53	<0.001	0.75 ± 0.05	<0.001	0.76 ± 0.06	<0.001	0.74 ± 0.06	<0.001
	ShallowConvNet [37]	77.42 ± 7.37	<0.001	0.76 ± 0.05	<0.001	0.73 ± 0.06	<0.001	0.75 ± 0.06	<0.001
	EEGNet [45]	77.79 ± 8.79	<0.001	0.76 ± 0.07	<0.001	0.75 ± 0.08	<0.001	0.76 ± 0.07	<0.001
	LSTM [30]	60.61 ± 5.58	<0.001	0.60 ± 0.04	<0.001	0.60 ± 0.03	<0.001	0.61 ± 0.03	<0.001
	TSCNN *(Ours)*	84.40 ± 8.03	-	0.85 ± 0.07	-	0.85 ± 0.07	-	0.85 ± 0.07	-

* p-value when comparing accuracies of TSCNN with other methods

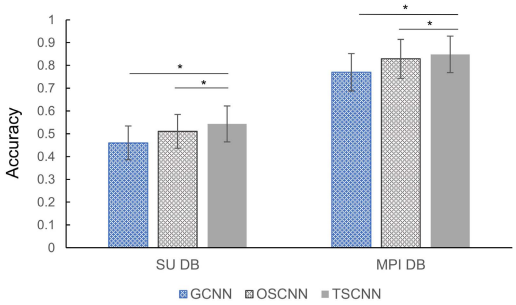

Fig. 9. Average classification accuracies over ten-folds and all subjects using OSCNN, GCNN, and TSCNN for SU DB (6-class semantic category) and MPI DB (2-class semantic category). Significant differences between accuracies from the different networks are shown (* $p < 0.05$ with Bonferroni correction, error bars: standard deviation across subjects).

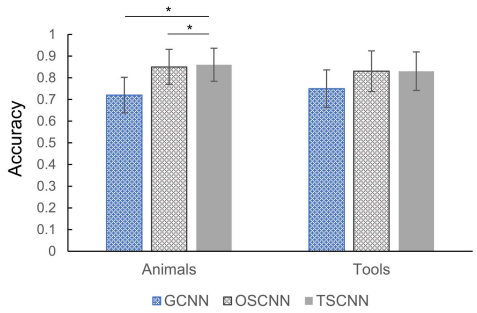

Fig. 11. Average classification accuracies per class over ten-folds and all subjects using OSCNN, GCNN, and TSCNN for MPI DB. Significant differences between accuracies from the different networks are shown (* $p < 0.05$ with Bonferroni correction, error bars: standard deviation across subjects).

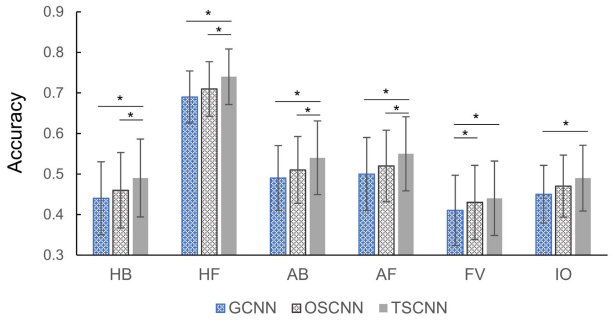

Fig. 10. Average classification accuracies per class over ten-folds and all subjects using OSCNN, GCNN, and TSCNN for SU DB. Significant differences between accuracies from the different networks are shown (* $p < 0.05$, error bars: standard deviation across subjects).

significantly higher accuracies for both datasets (6-class: 54.28 ± 7.89%, and 2-class: 84.40 ± 8.03%).

Fig. 10 shows the statistical comparison between accuracies for each class of SU DB. TSCNN showed significantly higher accuracies than the other two methods for all classes. Comparison between accuracies of each class was also performed for MPI DB. TSCNN obtained higher or similar accuracies for all classes. Statistical results revealed; significantly higher accuracies than the other methods for animal class; however, there were no significant differences between the accuracies for tool class (see Fig. 11).

E. Comparison With State-of-the-art Methods

We compared the classification accuracy, precision, sensitivity, and specificity of the proposed method with the state-of-the-art methods (see Table V). We implemented a baseline method proposed by Kaneshiro [1] that uses ERP as feature extraction and LDA as the classifier. ShallowConvNet [37] and EEGNet [45] have been used before in multiple EEG paradigms and have shown outstanding performance. Therefore, they have become a state-of-the-art method when classifying EEG signals. ShallowConvNet consists of two CNN layers, a mean pooling layer, and a fully connected layer with softmax activation for classification. EEGNet consists of three CNN layers, with max-pooling layers and a fully connected layer with softmax activation for classification. ShallowConvNet and EEGNet networks were fit using Adam optimizer and minimizing cross-entropy loss function. For both CNN-based networks, we used the same number of layers proposed in the original papers as well as the pooling layers, however, the kernel size was set to different values. We reported the highest obtained results. Additionally, we evaluated an LSTM network with the Swish activation function and Bagging theory proposed by Zheng et al. [30]. We set different hidden sizes (128 or 256) and the number of layers (1 or 2). Similar to the original implementation, we obtained better results when setting the hidden size to 128 and using one layer. This work also attempted to classify EEG signals evoked during the

presentation of visual stimuli, as a result, we consider suitable its implementation and comparison with the proposed method. Results of the state-of-the-art methods were compared with the proposed method using paired t-test analysis, this was done to determine significant differences and evaluate the relevance of the proposed network. The significance level was set at $p = 0.05$ with Bonferroni correction.

To ensure a fair comparison between the proposed method and other methods, we implemented the state-of-the-art methods by following the description in each of the published papers and performing a parameter search; additionally, we set the seed to the same value for all experiments, this ensures that the trials used on training and testing set were the same for all experiments. Table V shows the comparison and statistical results of the proposed methods and other methods. For both datasets, the lowest classification accuracies were obtained using the LSTM as a classification method [30] (SU DB: 38.06 $\pm$ 1.88%; MPI DB: 60.61 $\pm$ 5.58%). The individual classification performances for SU DB and MPI DB are shown in Supplementary Table V and Table VI, respectively. TSCNN outperformed the other methods; moreover, statistical analysis revealed significant differences between TSCNN and all other methods. Comparison between the averaged computation time per subject was conducted (see Supplementary Table VII). The proposed method required the highest computational time, followed by the GCNN. CNN-based methods had the lowest computational time for both datasets. We generated t-sne plots to compare the feature extraction ability of the proposed method with the state-of-the-art methods, w(see supplementary Fig. 4 and Fig. 5). The proposed method showed better feature separability compared with the rest.

V. Discussion

This study attempted to classify the presented images into their semantic category (6-class and 2-class) using EEG signals. We found differences in the event-related potential analysis as well as in the functional connectivity analysis. These prove their potential to be used for semantic category classification of observed stimuli. As a result, our proposed methods used both and integrated them using a two-stream CNN. Our results significantly outperformed the state-of-the-art methods.

A. Differences in Event-Related Potentials

EEG analysis results revealed significant differences between semantic categories. For SU DB, we found a significant difference in the frontal, central, and occipito-temporal regions at 190 ms and in the frontal area at 300 ms between semantic categories. For MPI DB, we found significant differences in the central, temporal, and occipital regions at different time points. Previous studies have explored brain responses to semantic categories and found significant differences depending on the semantic category of the stimulus shown to the subjects. In this regard, Haxby et al. [10] concluded that there is a specific region of the brain dedicated to processing faces. Therefore, we can conclude that different semantic categories could activate brain-dedicated regions.

B. Differences in Functional Connectivity

We calculated wPLI as functional connectivity to explore the difference during each semantic stimulus. Previous studies had shown that objects and semantic categories recognition are processed in the occipital and temporal region of the brain [10]. Moreover, these agree with the two-stream hypothesis; which is a neural model for the processing of human vision. It defines the ventral-temporal and dorsal-parietal streams (what and where pathways) that process information regarding object features [46]. Additionally, a study showed that the activation of occipital and temporal regions and connectivity between occipito-temporal and frontal areas are present in lexical tasks. These connectivities are known to play a major role in lexical-semantic language processes [47]. Our results are in congruence with those findings since connectivity was mainly present between frontal and occipital regions. However, we also showed that functional connectivity differs between the semantic categories for both datasets. This supports its use for classifying EEG signals into semantic categories.

C. Comparisons of Classification Performance

Our results demonstrated that it is possible to classify EEG signals elicited during the presentation of stimuli with higher accuracy compared to other methods. We integrated channel-wise features and functional connectivity using TSCNN; and obtained significantly higher results than other methods. The analyzed datasets had different pre-processing methods, which can represent a limitation when performing classification. Therefore, we decided to analyze each dataset separately and select optimal hyper-parameters using grid search.

When using only GCNN accuracies were higher than the chance level, which shows that GCNN extracts relevant features when classifying semantic categories. Additionally, we analyzed the influence of the electrodes' distance and functional connectivity, by constructing the graph using each one separately or both. We obtained higher classification performance when using both values, therefore we can infer that GCNN can simulate brain connections (local and distant) through electrode distance and functional connectivity and at the same time increase classification performance. Even though results were higher when using just electrode distance than when using just functional connectivity, there was no significant difference, showing that both contributed similarly to the final performance.

GCNN results were outperformed by the proposed method. This supports our initial hypothesis; that even though GCNN can simulate brain connections (local and distant); channel-wise features could be ignored. We notice that although ShallowConvNet [37] and EEGNet [45] use convolutional neural networks for classification same as OSCNN; OSCNN obtained higher results; which we attribute to the selection of parameters such as the number of layers, kernel size, and pooling layers. LSTM-based network obtained the lowest accuracies, showing that convolution operations are more adequate for classifying object perception.

For SU DB, the confusion matrix showed that HF class is the most distinctive category. As mentioned previously, multiple studies have reported distinctive brain patterns for face perception [10], [14], which can explain the high accuracy obtained for this semantic category. At the same time, FV class was the least distinctive category, which is in accordance with previous studies [1]. The comparison of classification accuracies of each class between the TSCNN with GCNN and OSCNN for all subjects revealed that accuracies increased for most of the classes, and remained similar for FV class. Moreover, FV and IO classes had the lowest accuracies and the model was confused between these two classes. In this regard, FV and IO classes can be grouped as objects with no motion and the remaining classes as animated objects, which can explain the above statement. For MPI DB, the confusion matrix when using TSCNN showed that the animal class is more distinguishable than the tool class. Similarly, when comparing the accuracies of each class using TSCNN with GCNN and OSCNN, significant differences were found only in the animal class. The animal class benefits more than the tool class from our proposed architecture, previous studies have shown the different brain patterns present when decoding animal and object [48]. Since we simulated local and distant brain connections, we could assume that our network decoded the brain patterns corresponding to the animal class better than the other networks, meanwhile, tools class is decoded with similar performance for all methods.

D. Limitation

One limitation of this study is the use of the CNN models, which are considered black-box [49] since it is not clear which are the main features extracted for the classification. We assume that the CNN extracts time channel-wise temporal features since we applied a kernel on the temporal dimension to the signal (1, K). Although the use of CNN increased the classification performance, we agree that more research needs to be performed to confirm this hypothesis. wPLI is used for calculating the functional connectivity between channels, however, this requires a high density of EEG channels. Therefore, this represents a limitation in relationship with other state-of-the-art methods, which can be applied to low-density EEG data. Moreover, channels' information is needed to calculate the distance between electrodes; some of the public datasets do not include this information as a result the proposed method can not be applied.

Finally, due to the pre-processed methods used in [1] and [6] there is a limitation on the analysis we could perform. In future works, we decided to recruit subjects and perform EEG experiments.

VI. Conclusion

We investigated the differences in brain signals in the temporal domain and functional connectivity values between semantic categories. This revealed that differences occurred at different locations and time points and that the relation between channels varies depending on the semantic category analyzed. Therefore, we assume that training a GCNN could simulate brain connectivity and extract relevant features, however, channel-wise features could be ignored. As a result, we proposed the TSCNN that uses GCNN and a CNN to take advantage of functional connectivity and extract channel-wise features. TSCNN exhibited significantly higher accuracy than other state-of-the-art methods and ablations studies. This supports our assumption and proves that our method is relevant.

We decided to further explore different classification methods for improving accuracy. Even though there is not enough evidence of differences in EEG rhythms for different semantic categories, we decided to further analyze this possibility and whether fatigue due to a long time of the experiments can influence the results. As mentioned before, discriminate information related to object recognition could be found at 100 ms, making this a rapid process, when compared to other BCI paradigms such as motor imagery [28], this opens the possibility of applying this paradigm to fast and reliable BCI systems reducing at the same time fatigue due to long exposure of different stimulus [12]. BCI controls usually do not have a connection with the semantics of the task, this could affect drastically the performance [50], object recognition avoids this problem, as a result, better results could be obtained. Additionally, the proposed method could be applied to other intuitive BCI paradigms such as visual imagery and speech imagery since provided stimuli could also be divided by their semantic categories. This could improve the interaction with humans who, due to various reasons, are unable to communicate using conventional methods.

References

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

[2] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as Bayesian inference," *Annu. Rev. Psychol.*, vol. 55, no. 1, pp. 271–304, 2004.

[3] D. R. Vogel et al., *Persuasion and the Role of Visual Presentation Support: The UM/3M Study*. CA, USA: Management Information Systems Research Center, School of Management, 1986.

[4] L. Ko, O. Komarov, and S.-C. Lin, "Enhancing the hybrid BCI performance with the common frequency pattern in dual-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 7, pp. 1360–1369, Jul. 2019.

[5] R. M. Cichy, D. Pantazis, and A. Oliva, "Resolving human object recognition in space and time," *Nature Neurosci.*, vol. 17, no. 3, pp. 455–462, Jan. 2014.

[6] I. Simanova, M. van Gerven, R. Oostenveld, and P. Hagoort, "Identifying object categories from event-related EEG: Toward decoding of conceptual representations," *PLoS ONE*, vol. 5, no. 12, Dec. 2010, Art. no. e14465.

[7] K. Grill-Spector, "The neural basis of object perception," *Current Opinion Neurobiol.*, vol. 13, no. 2, pp. 159–166, Apr. 2003.

[8] C. B. Hart and W. J. Rose, "Visual feature extraction from voxel-weighted averaging of stimulus images in 2 fMRI studies," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 11, pp. 3124–3130, Nov. 2013.

[9] C. Wang, S. Xiong, X. Hu, L. Yao, and J. Zhang, "Combining features from ERP components in single-trial EEG for discriminating four-category visual objects," *J. Neural Eng.*, vol. 9, no. 5, Oct. 2012, Art. no. 056013.

[10] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, Sep. 2001.

[11] S. J. Hanson, T. Matsuka, and J. V. Haxby, "Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a 'face' area?" *NeuroImage*, vol. 23, no. 1, pp. 156–166, Sep. 2004.

[12] F. Putze et al., "Hybrid fNIRS-EEG based classification of auditory and visual perception processes," *Frontiers Neurosci.*, vol. 8, p. 373, Nov. 2014.

[13] J. Kalafatovich and M. Lee, "Subject-independent object classification based on convolutional neural network from EEG signals," in *Proc. 9th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2021, pp. 1–4.

[14] R. J. Itier, "N170 or N1? Spatiotemporal differences between object and face processing using ERPs," *Cerebral Cortex*, vol. 14, no. 2, pp. 132–142, Feb. 2004.

[15] S. Bentin, T. Allison, A. Puce, E. Perez, and G. McCarthy, "Electrophysiological studies of face perception in humans," *J. Cognit. Neurosci.*, vol. 8, no. 6, pp. 551–565, 1996.

[16] M. G. Philiastides, "Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram," *J. Neurosci.*, vol. 26, no. 35, pp. 8965–8975, Aug. 2006.

[17] D. S. Bassett and O. Sporns, "Network neuroscience," *Nature Neurosci.*, vol. 20, no. 3, pp. 353–364, Feb. 2017.

[18] B. Sun, H. Zhang, Z. Wu, Y. Zhang, and T. Li, "Adaptive spatiotemporal graph convolutional networks for motor imagery classification," *IEEE Signal Process. Lett.*, vol. 28, pp. 219–223, 2021.

[19] Z. Wang, Y. Tong, and X. Heng, "Phase-locking value based graph convolutional neural networks for emotion recognition," *IEEE Access*, vol. 7, pp. 93711–93722, 2019.

[20] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul./Sep. 2020.

[21] D. Zhang, K. Chen, D. Jian, and L. Yao, "Motor imagery classification via temporal attention cues of graph embedded EEG signals," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 9, pp. 2570–2579, Sep. 2020.

[22] C.-R. Phang, F. Noman, H. Hussain, C.-M. Ting, and H. Ombao, "A multi-domain connectome convolutional neural network for identifying schizophrenia from EEG connectivity patterns," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1333–1343, May 2020.

[23] M. Vinck, R. Oostenveld, M. Van Wingerden, F. Battaglia, and C. M. A. Pennartz, "An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias," *Neuroimage*, vol. 55, no. 4, pp. 1548–1565, Apr. 2011.

[24] K. Yoshinaga et al., "Comparison of phase synchronization measures for identifying stimulus-induced functional connectivity in human magnetoencephalographic and simulated data," *Frontiers Neurosci.*, vol. 14, p. 648, 2020.

[25] M. Lee et al., "Network properties in transitions of consciousness during propofol-induced sedation," *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, 2017.

[26] C. Du, C. Du, L. Huang, H. Wang, and H. He, "Structured neural decoding with multitask transfer learning of deep neural network representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 600–614, Feb. 2022.

[27] K. Han et al., "Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex," *NeuroImage*, vol. 198, pp. 125–136, Sep. 2019.

[28] K. Seeliger et al., "Convolutional neural network-based encoding and decoding of visual object recognition in space and time," *NeuroImage*, vol. 180, pp. 253–266, Oct. 2018.

[29] T. Carlson, D. A. Tovar, A. Alink, and N. Kriegeskorte, "Representational dynamics of object vision: The first 1000 MS," *J. Vis.*, vol. 13, no. 10, pp. 1–19, Aug. 2013.

[30] X. Zheng, W. Chen, Y. You, Y. Jiang, M. Li, and T. Zhang, "Ensemble deep learning for automated visual classification using EEG signals," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107147.

[31] X. Zheng and W. Chen, "An attention-based bi-LSTM method for visual object classification via EEG," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102174.

[32] H. Ahmed, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, "Object classification from randomized EEG trials," 2020, *arXiv:2004.06046*.

[33] J. Kalafatovich, M. Lee, and S.-W. Lee, "Decoding visual recognition of objects from EEG signals based on attention-driven convolutional neural network," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 2985–2990.

[34] E. L. Hall, S. E. Robson, P. G. Morris, and M. J. Brookes, "The relationship between MEG and fMRI," *NeuroImage*, vol. 102, pp. 80–91, Nov. 2014.

[35] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI," *J. Neural Eng.*, vol. 16, no. 2, Apr. 2019, Art. no. 026007.

[36] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli, and W. Geng, "Surface-electromyography-based gesture recognition by multi-view deep learning," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2964–2973, Oct. 2019.

[37] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, Mar. 2017.

[38] S. Bagchi and D. R. Bathula, "EEG-ConvTransformer for single-trial EEG-based visual stimulus classification," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108757.

[39] J. Jin et al., "A novel classification framework using the graph representations of electroencephalogram for motor imagery based brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 20–29, 2022.

[40] Y. Hou et al., "GCNs-net: A graph convolutional neural network approach for decoding time-resolved EEG motor imagery signals," 2020, *arXiv:2006.08924*.

[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 448–456.

[42] L. S. Imperatori et al., "EEG functional connectivity metrics wPLI and wSMI account for distinct types of brain functional interactions," *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, Dec. 2019.

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

[44] T. Gandhi, B. K. Panigrahi, and S. Anand, "A comparative study of wavelet families for EEG signal classification," *Neurocomputing*, vol. 74, no. 17, pp. 3051–3057, 2011.

[45] V. Lawhern, A. Solon, N. Waytowich, S. M. Gordon, C. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, p. 056013, 2018.

[46] J. V. Haxby et al., "Dissociation of object and spatial visual processing pathways in human extrastriate cortex," *Proc. Nat. Acad. Sci. USA*, vol. 88, no. 5, pp. 1621–1625, 1991.

[47] A. Ewald, S. Aristei, G. Nolte, and R. A. Rahman, "Brain oscillations and functional connectivity during overt language production," *Frontiers Psychol.*, vol. 3, p. 166, 2012.

[48] Z. Cao et al., "Distinct brain activity in processing negative pictures of animals and objects—The role of human contexts," *NeuroImage*, vol. 84, pp. 901–910, Jan. 2014.

[49] Z. Zeng, C. Miao, C. Leung, and J. J. Chin, "Building more explainable artificial intelligence with argumentation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2018, pp. 8044–8045.

[50] N. Kosmyna, J. T. Lindgren, and A. Lécuyer, "Attending to visual stimuli versus performing visual imagery as a control strategy for EEG-based brain-computer interfaces," *Sci. Rep.*, vol. 8, no. 1, pp. 1–14, Dec. 2018.