Psychoradiology, 2024, **4**, kkae021

DOI: 10.1093/psyrad/kkae021

Advance access publication date: 4 November 2024

Research Article

Explorations of using a convolutional neural network to understand brain activations during movie watching

Wonbum Sohn[1,2], Xin Di[1,*], Zhen Liang[3], Zhiguo Zhang[4] and Bharat B. Biswal[1,*]

[1]Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA
[2]Rutgers Biomedical and Health Sciences, Rutgers School of Graduate Studies, Newark, NJ 07039, USA
[3]School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, 51806, China
[4]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, 518060, China
*Correspondence: Xin Di, xin.di@njit.edu; Bharat B. Biswal, bharat.biswal@njit.edu

Abstract

Background: Naturalistic stimuli, such as videos, can elicit complex brain activations. However, the intricate nature of these stimuli makes it challenging to attribute specific brain functions to the resulting activations, particularly for higher-level processes such as social interactions.

Objective: We hypothesized that activations in different layers of a convolutional neural network (VGG-16) would correspond to varying levels of brain activation, reflecting the brain's visual processing hierarchy. Additionally, we aimed to explore which brain regions would be linked to the deeper layers of the network.

Methods: This study analyzed functional MRI data from participants watching a cartoon video. Using a pre-trained VGG-16 convolutional neural network, we mapped hierarchical features of the video to different levels of brain activation. Activation maps from various kernels and layers were extracted from video frames, and the time series of average activation patterns for each kernel were used in a voxel-wise model to examine brain responses.

Results: Lower layers of the network were primarily associated with activations in lower visual regions, although some kernels also unexpectedly showed associations with the posterior cingulate cortex. Deeper layers were linked to more anterior and lateral regions of the visual cortex, as well as the supramarginal gyrus.

Conclusions: This analysis demonstrated both the potential and limitations of using convolutional neural networks to connect video content with brain functions, providing valuable insights into how different brain regions respond to varying levels of visual processing.

Keywords: convolutional neural network; deep learning; default mode network; lateral occipital complex; naturalistic condition; supramarginal gyrus; visual cortex

Introduction

Recently, naturalistic stimuli, such as movies and stories, have increasingly been employed to study brain functions in human neuroimaging research. This approach offers several advantages over traditional task-based functional magnetic resonance imaging (fMRI) experiments. One of the primary benefits is that naturalistic stimuli closely resemble real-life situations, enabling the elicitation of complex cognitive processes. On the other hand, compared with resting-state fMRI, naturalistic stimuli allow for a higher level of experimental control, resulting in improved participant cooperation and increased reliability of research findings. A pivotal study by Hasson *et al.* demonstrated that different participants watching the same video stimulus exhibited similar patterns of brain activity across distributed brain regions (Hasson *et al.*, 2004). This finding led to the widespread use of inter-participant correlation as a means to identify activity and connectivity patterns associated with various stimuli (Nastase, 2019; Chen *et al.*, 2020; Di and Biswal, 2020). Despite these advancements, one major challenge is linking the observed brain data to the contents of naturalistic stimuli, such as videos and audios, due to their inherent complexity.

Numerous analytical approaches have emerged to study the complexities of naturalistic stimuli. One conventional method is utilizing human subjective ratings. For instance, researchers have asked participants to rate their perceived motion while watching a cartoon video and then employed general linear model to map brain responses related to motion perception. This approach identified motion-sensitive brain regions in the middle temporal lobe (Rao *et al.*, 2007). Furthermore, subjective affective states can be reported and linked to brain activations and dynamic connectivity (Raz *et al.*, 2012; Sun *et al.*, 2022). Another approach is manually tagging objects of interest to investigate category-specific brain activations (Richardson *et al.*, 2018). Advancements in machine learning technologies have also been leveraged. Previous studies used traditional computer vision models to extract global motion, local motion, and residual models based on motion flow and patterns from videos. They found that the medial posterior parietal cortex, V5+, and V1–V4 were activated in the scenes of the global motion model, local motion model, and residual model, respectively (Bartels *et al.*, 2008). Celik *et al.* have built encoding models of various objects (car, bridge, etc.) from video stimuli to study category representations in the cerebral cortex (Çelik *et al.*, 2021).

Received: 19 July 2024; **Revised:** 9 September 2024; **Accepted:** 31 October 2024

Recently, convolutional neural networks (CNNs) have been used to extract visual features of videos, particularly in the context of face processing (Jiahui *et al.*, 2022; Hu *et al.*, 2023).

The naturalistic stimuli have been selected to explore intricate social functions such as the theory of mind and empathy (Richardson *et al.*, 2018). However, the field still lacks machine learning models that can effectively describe various aspects of social functions due to the complexity of the naturalistic stimuli. A recent study by McMahon and colleagues employed multiple machine learning models to extract different levels of features from videos containing social interactions (McMahon *et al.*, 2023). They established a hierarchy of social interactions, primarily linked to the temporal lobe regions. Nevertheless, the higher-level features in their hierarchy still rely on manually selected features. In our current study, we aim to test the hypothesis of whether we can extract features related to social functions using CNNs. We systematically investigate how different convolutional layers are associated with the hierarchy of various brain regions. This approach may offer valuable insights into understanding the neural basis of social interactions and potentially uncover novel findings that were previously limited by manual feature selection methods.

CNNs have demonstrated remarkable success in computer vision (Simonyan and Zisserman, 2015; Krizhevsky *et al.*, 2017). One of fundamental elements in CNNs is the convolutional kernels, which extract local features from data. As the data progresses through deeper layers of convolutional kernels, more complex features are extracted. While CNNs are typically trained on large-scale image datasets for image recognition, encompassing 1000 categories (Deng *et al.*, 2009), we posit that a CNN trained on the ImageNet dataset might have learned information relevant to social interactions. To explore this hypothesis, we investigate how features extracted from various kernels of convolutional layers correlate with brain activations in different brain regions. In a recent study, Hu and colleagues utilized a pre-trained VGG-16 CNN to extract features from different layers while analyzing affective videos (Hu *et al.*, 2023). They discovered that brain microstates calculated from electroencephalogram data were only correlated with features from deeper convolutional layers (layers 11, 12, and 13). Building on this work, our current study employs fMRI data, which provides superior spatial resolution. This enables us to examine how different brain regions are associated with the features extracted from diverse kernels of convolutional layers. By leveraging the strengths of fMRI, we aim to gain deeper insights into the relationship between neural activations and the hierarchical visual representations generated by CNNs.

In this study, we employed a single pre-trained VGG-16 network, one of image feature extractors, to analyze a short, animated movie and extract features at different levels from the convolutional layers. Our primary aim was to investigate how and where these diverse levels of features are represented in the human brain. To accomplish this, we analyzed fMRI data from young adult volunteers while they watched the same movie clip. We utilized a generalized linear model approach to map brain regions whose temporal activity pattern matched the feature activity pattern from specific kernels of a convolutional layer. Our hypothesis revolves around the notion that brain activation patterns will exhibit a hierarchy from low-level visual areas to high-level areas related to social interaction and empathy through CNN's hierarchical feature maps. Of particular interest to us were the brain regions associated with higher convolutional layers. Considering that VGG-16 was trained for image classification, we postulated that higher convolutional layers might primarily represent categorical features, possibly linking to the ventral visual pathway in the brain. In addition, due to the rich image context in the movie, it could also encompass action, motion, and social information, which may be more prevalent in deeper convolutional layers of VGG-16. We were especially intrigued by whether the dorsal visual pathway, and even the third "social" pathway, could also be associated with features from deeper layers. If this proves to be true, VGG-16 could serve as a valuable tool for studying social functions using video stimuli, extending beyond its conventional image classification capabilities.

Materials and Methods

fMRI Data

The fMRI data were obtained from openneuro (https://openneuro.org/; accession no. ds000228). Only adults' data ($n = 33$) were used for this study. The effective sample consisted of 17 females and 12 males (age range 18–39 years, mean = 24.6, SD = 5.3), excluding participants with excessive head motion or poor coverage according to the criteria mentioned in the previous study (Di and Biswal, 2020).

During the fMRI acquisition, all participants watched the "Partly Cloudy" animation by Pixar (https://www.pixar.com/partly-cloudy#par-tly-cloudy-1) for ∼6 minutes with a black screen of first 10 seconds (1–5 TRs) (Richardson *et al.*, 2018). Structural and functional MRI data were acquired by a 3 Tesla Siemens Tim Trio scanner at the Massachusetts Institute of Technology. The T1-weighted structural MRI data were gathered in 176 interleaved sagittal slices with 1mm isotropic. All fMRI data were acquired by a gradient-echo EPI (repetition time (TR) = 2 s, echo time (TE) = 30 ms, flip angle = 90°). A total of 168 fMRI images were measured from each participant.

The fMRI data were analyzed using SPM12 and MATLAB (R2021b) as described in our previous study (Di and Biswal, 2020). Briefly, first, the anatomical T1 images of all participants were segmented into six segments. Afterward, all skulls were removed from the T1 images. All functional images were realigned concerning the first image. In this process, the degree of translation and rotation was calculated, and participants with a maximum framewise displacement >1.5 mm or 1.5° were removed (Di and Biswal, 2015). The remaining functional images were coregistered with the skull-stripped anatomical image of the same participant. Next, the anatomical and functional images were normalized to MNI space, and the voxel size was resampled to $3 \times 3 \times 3$ mm^3. Finally, the functional images were spatially smoothed through an 8-mm Gaussian kernel.

Video Analysis

Because brain activations occurring within each TR are represented in a single fMRI image, the video content needs to be grouped into TR units to enable comparison between the continuous video content and the fMRI data. Given the TR of 2 seconds, each fMRI image corresponds to 2 seconds of video contents. Therefore, we averaged the video frames for every 2 seconds before extracting features from the video. Since the frame rate was 24 frames/s, 48 frames were averaged into one image. As a result, the 8370 total frames of the video clip were converted to 175 images.

To examine the correlation between fMRI images generated at each TR and video features, we utilized CNNs, which are specialized for image analysis, to process the image-based videos. Specifically, similar to studies exploring the relationship between

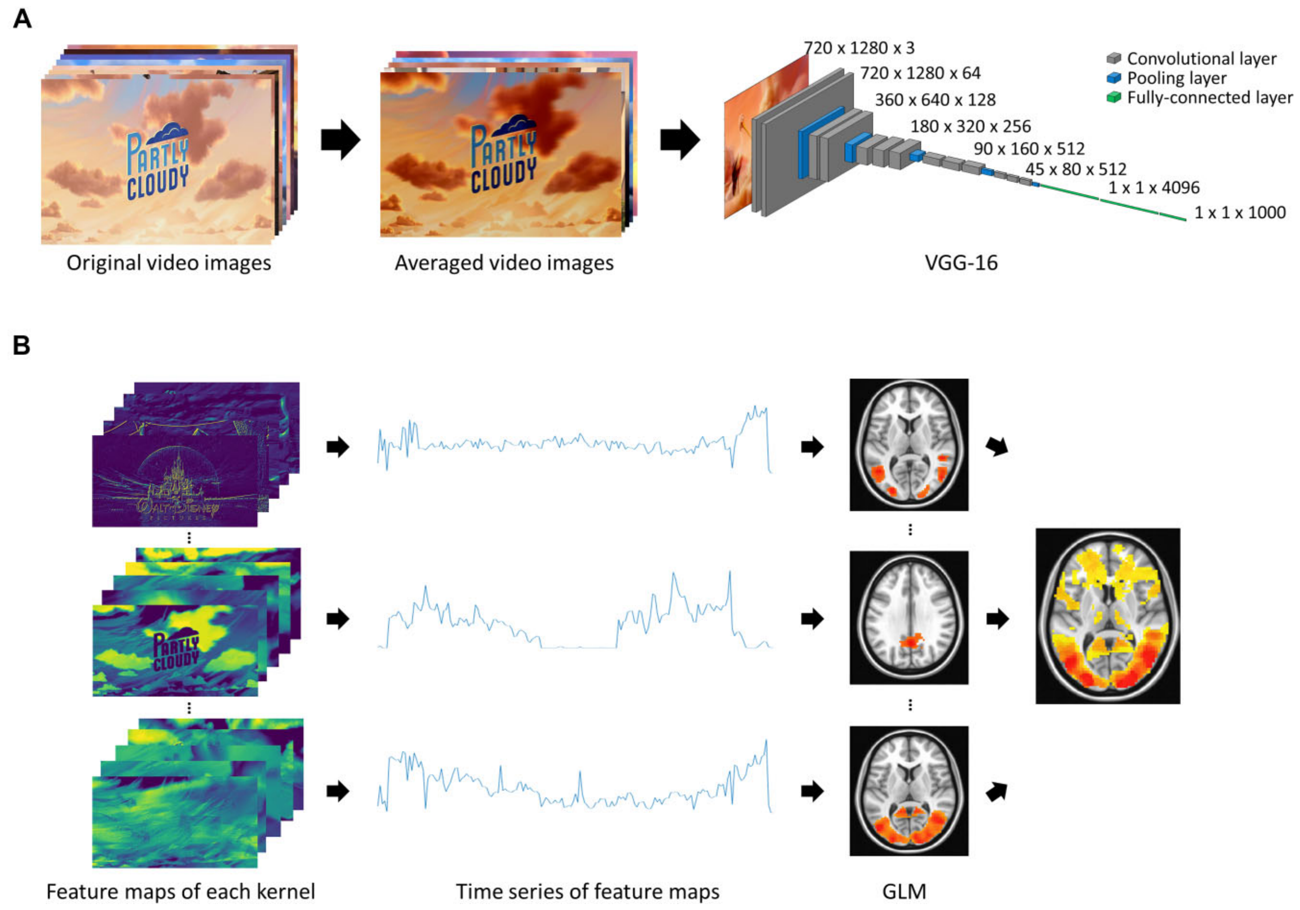

Figure 1: Illustration of video feature extraction and general linear model (GLM) analysis. (**A**) 8370 images of the original video were converted into 175 images based on TR 2 s and the 175 images were used in VGG-16 for feature extraction. (**B**) 2D feature maps from each kernel were transformed into time series, and then the time series was used for GLM on the fMRI data. Kernel activation maps were calculated by averaging the activation maps of every kernel of a convolutional layer in VGG-16.

emotion dynamics and stimuli, we employed the pre-trained VGG-16 model to analyze the image data. VGG-16 consists of 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers (Fig. 1A). Each averaged image with 720 × 1280 × 3 dimensions was fed to the VGG-16 model. The input images were convolved with 64 3 × 3 × 3 kernels in the first convolutional layer and then generated 720 × 1280 × 64 feature maps. Feature maps show the location and intensity of various patterns contained in the input. These were used as an input of the second convolutional layer and convolved with 64 3 × 3 × 64 kernels, generating 720 × 1280 × 64 feature maps. Afterward, the size of the feature maps was reduced to 360 × 640 × 64 through the max-pooling layer. This process has repeated five times, and finally, the size of the feature maps became 45 × 80 × 512. The last max-pooling layer and the three fully connected layers of VGG-16 were not used because this step aimed to extract feature maps from the averaged input images instead of classification. Each layer receives the output feature maps of the previous layer as input and generates higher-level feature maps. As a result, low-level features such as edges, lines, and similar color regions were detected in shallow convolutional layers. On the other hand, abstract and synthetic high-level features were shown in the deeper layers.

Linking VGG-16 Activations to Brain Activations

For each TR, each kernel of a convolutional layer of the VGG-16 model generates a distinct activation map. In the context of general linear model (GLM) analysis—which scrutinizes brain regions exhibiting patterns akin to a specific time series during the measurement period—each activation map was averaged into a single value. Consequently, with 175 input images sampling the entire video, a 1D time series of 175 time was generated for each kernel. This resulted in 64 time series for the first and second convolutional layers, and 512 time series for eleventh, twelfth, and thirteenth convolutional layers (Fig. 1B). We chose to use a crude measure of averaged activations to streamline the analysis, enabling us to investigate various kernels and layers. It is crucial to note that fMRI is also a coarse measure of neural activity, with each voxel representing over 600 000 neurons. While we were aware of more sophisticated methods like representational similarity analysis (Kriegeskorte *et al.*, 2008), we opted for the straightforward approach for this initial exploratory analysis.

To select two convolutional layers with a large difference between feature maps, we calculated the correlation between layers. We averaged the 1D time series data that exist as many as the number of kernels in each convolutional layer to obtain one averaged 1D time series data for each layer. We selected the first and thirteenth convolutional layers with the lowest correlation ($r = -0.054$) for GLM and reverse analysis.

We performed voxel-wise GLM using the 1D time series of the kernels and preprocessed fMRI data. First, the first five TRs of the fMRI data obtained from the black screen were excluded because they were unrelated to the video. Next, to match with fMRI data length, the 1D time series were used only for the first 163 TRs.

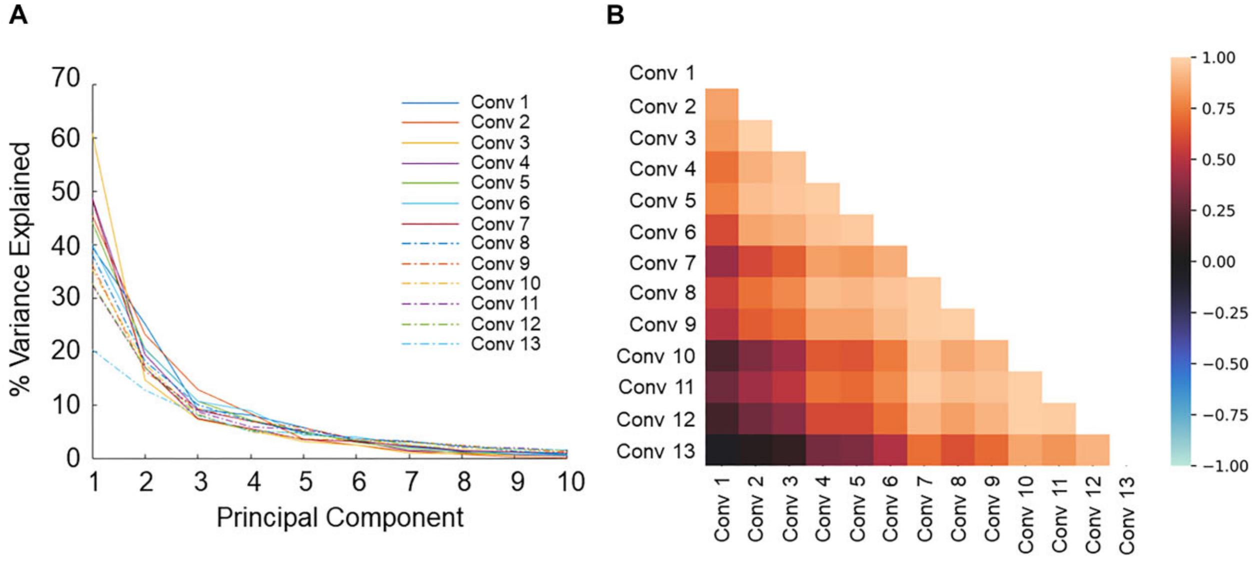

Figure 2: The characteristics of the convolutional layers. (**A**) The variance explained by the first 10 principal components for the 13 convolutional layers. The variance percentage of PC1∼2 became lower when the convolutional layers went deeper. (**B**) A correlation matrix of averaged time series across the 13 convolutional layers. The correlation coefficient was higher between adjacent layers and lower between distant layers.

Friston's 24-parameter model was also added as a regressor to consider the effect of head motion (Friston *et al.*, 1996). To check the brain regions activated by the 1D time series of video features, the contrast was set to 1 for only the video time series and 0 for the remaining Friston's 24-parameter model parts. Finally, a group-level one-sample *t*-test was performed for each kernel to confirm the brain networks commonly activated by the features across the participants. After that, the cluster of the GLM results was performed based on $P < 0.001$. Then, the group-level one-sample *t*-test results of all kernels for each convolutional layer were added to see which functional network of the brain was related as the convolutional layer deepened.

Analysis of Feature Maps (Reverse Analysis)

In the voxel-wise GLM analysis, we identified brain regions that were activated by features extracted from different layers of VGG-16. We then conducted a reverse analysis to identify patterns in the feature maps, providing further insights into which features might be associated with the observed brain activations. Specially, we were interested in the brain activations in the default mode network and visual cortex prominent from features of the first convolutional layer and in the supramarginal gyrus and lateral occipital complex regions prominent from features of the thirteenth convolutional layer.

In the first convolutional layer, the brain's functional network obtained from the GLM with the kernel 6 related to the default mode network, especially posterior cingulate cortex, was masked with 0 and 1 to create a posterior cingulate mask. The generated posterior cingulate mask was applied to the preprocessed fMRI data of 29 participants, leaving only the intensities of the brain regions related to the posterior cingulate cortex and converting the intensities of the other regions to 0. The intensities of the remaining posterior cingulate regions were averaged with one value for each TR time and then made into a time series with a length of 165 TRs for each participant. To determine the TR time at which the posterior cingulate region is actively activated, the time series ensemble of 29 participants was averaged into one averaged fMRI time series. We identified the TR times with peaks in the averaged fMRI time series. Considering that the delay until the BOLD (blood oxygen level-dependent) signal was generated after stimulation was 2 TR (4 seconds), feature maps of the kernel 6 corresponding to the TR time after subtracting two TRs from the identified TR time were extracted. We repeated this process for other kernels 19, 24, 42, 55, and 58 related to the posterior cingulate to extract related feature maps. We investigated whether common patterns exist in the extracted feature maps.

The same method was applied to kernels 7, 43, and 63 of the first convolutional layer related to the visual cortex. In the thirteenth convolutional layer, kernels 9, 43, and 128 related to the lateral occipital complex and kernels 42, 103, 168, 276, 428, and 510 associated with the supramarginal gyrus were confirmed for the reverse analysis. After extracting the relevant feature maps, we analyzed whether there were common patterns in the feature maps for each brain region.

Results

Variability of Video Time Series with Layer Depth

We examined the variability between the 1D time series data obtained from each kernel in the convolutional layers and the correlation between the time series data across the convolutional layers. Principal component (PC) analysis was used to maximize the variability of kernels in each convolutional layer. In the shallow layers (first to seventh convolutional layers), there was a huge difference in the proportion of PC1 representing the features variability depending on the layer (Fig. 2A). On the other hand, in the deep layers (eighth to twelfth convolutional layers), there was little difference in the variability proportion of PC1 according to the layer except for the thirteenth convolutional layer. Overall, as the layer deepened, the proportion of PC1∼2 in variability tended to decrease. This means that a greater variety of features were extracted in deeper convolutional layers. To see the correlation between layers, a total of 13 averaged 1D time series data were used to calculate correlation coefficients (Fig. 2B). Correlation coefficients between adjacent convolutional layers were high. On the other hand, the correlation coefficients were lower as the distance

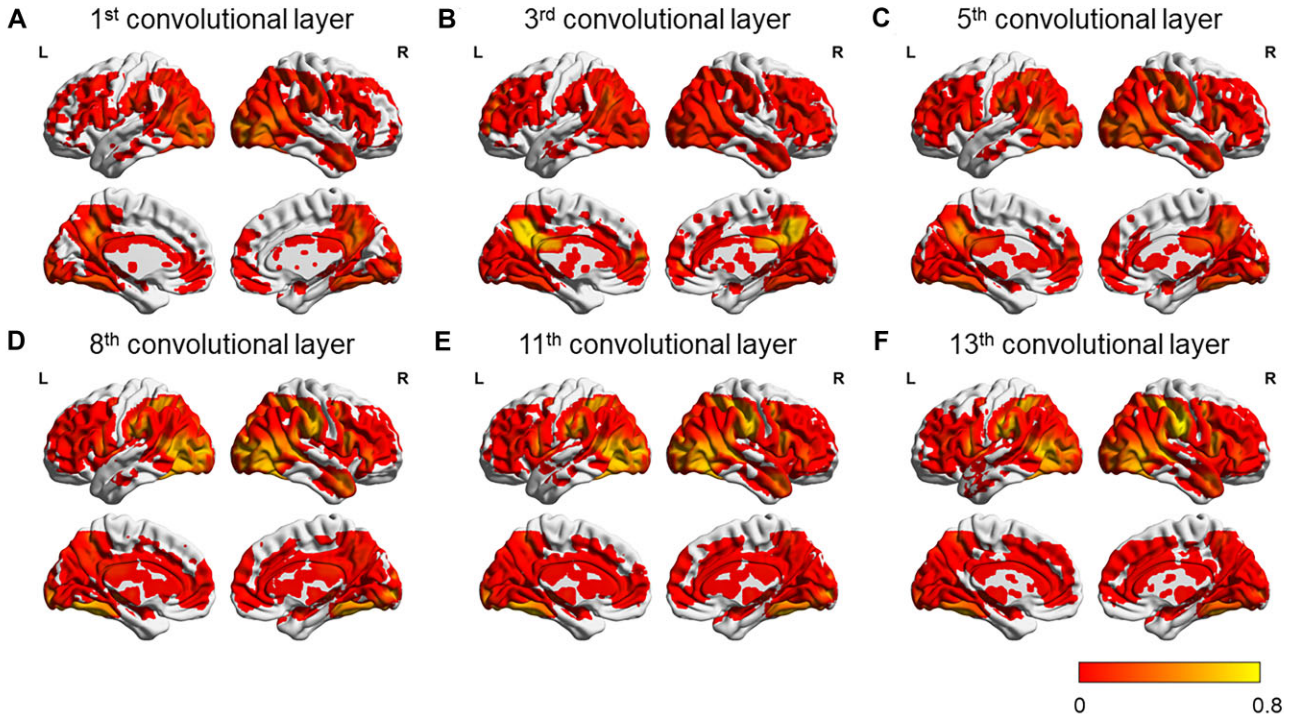

Figure 3: Activation probability by feature activations of the video in different kernels in the first (**A**), third (**B**), fifth (**C**), eighth (**D**), eleventh (**E**), and thirteenth (**F**) convolutional layers in VGG-16. Binary activation maps for each kernel were averaged within a layer to form the probability map, resulting in a range of 0 to 1.

of the convolutional layers increased. This was because CNN receives the output of the previous convolutional layer as an input and performs a convolution operation.

Voxel-Wise Brain Activation Results

After obtaining time series for each kernel in different convolutional layers, we performed voxel-wise analysis with brain activations measured with fMRI. Most of the GLMs produced statistically significant activations in various brain regions. We examined the distributions of activations in the brain in different convolutional layers of VGG-16 (Fig. 3). There were widespread activations associated with different kernels in all the layers. The visual cortex was more likely to be activated in all the convolutional layers. But the activation patterns outside the visual cortex conveyed a shift in different convolutional layers. Specifically, the posterior cingulate cortex was more likely to be activated in shallow layers such as the first and third convolutional layers. As layers went deeper, bilateral temporal and parietal regions were more likely to be activated.

We contrasted the brain activation probability maps between the shallowest layer (first convolutional layer) and the deepest layer (thirteenth convolutional layer). Figure 4 clearly shows the different activation distributions between the two layers. The first convolutional layer was more likely to be associated with the posterior visual cortex, as well as the posterior cingulate cortex, bilateral angular gyrus, and medial prefrontal cortex, which formed the default mode network. By contrast, the thirteenth convolutional layer was more likely to be associated with supramarginal gyrus, lateral occipital complex, and superior temporal sulcus.

Analysis of Feature Activation Patterns in VGG-16

To gain insights into the relationship between video features and brain activations across different regions, we conducted a reverse analysis. As a specific case, we focused on the brain regions that exhibited a higher propensity for activation based on features derived from the first and thirteenth convolutional layers. Specifically, we examined the visual cortex and posterior cingulate cortex region, which demonstrated a greater likelihood of activation in response to features from the first convolutional layer (Fig. 5). Additionally, we investigated the supramarginal gyrus and lateral occipital complex region, which exhibited a heightened probability of activation in response to features from the thirteenth convolutional layer (Fig. 6). To investigate independent relationships

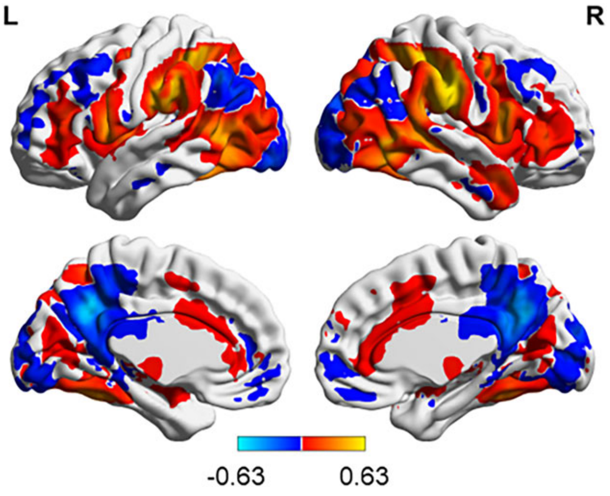

Figure 4: Differences in activation probability by feature activations in different kernels between the thirteenth and the first layers. Red is more active brain regions in the thirteenth convolutional layer, and blue is more active brain regions in the first convolutional layer.

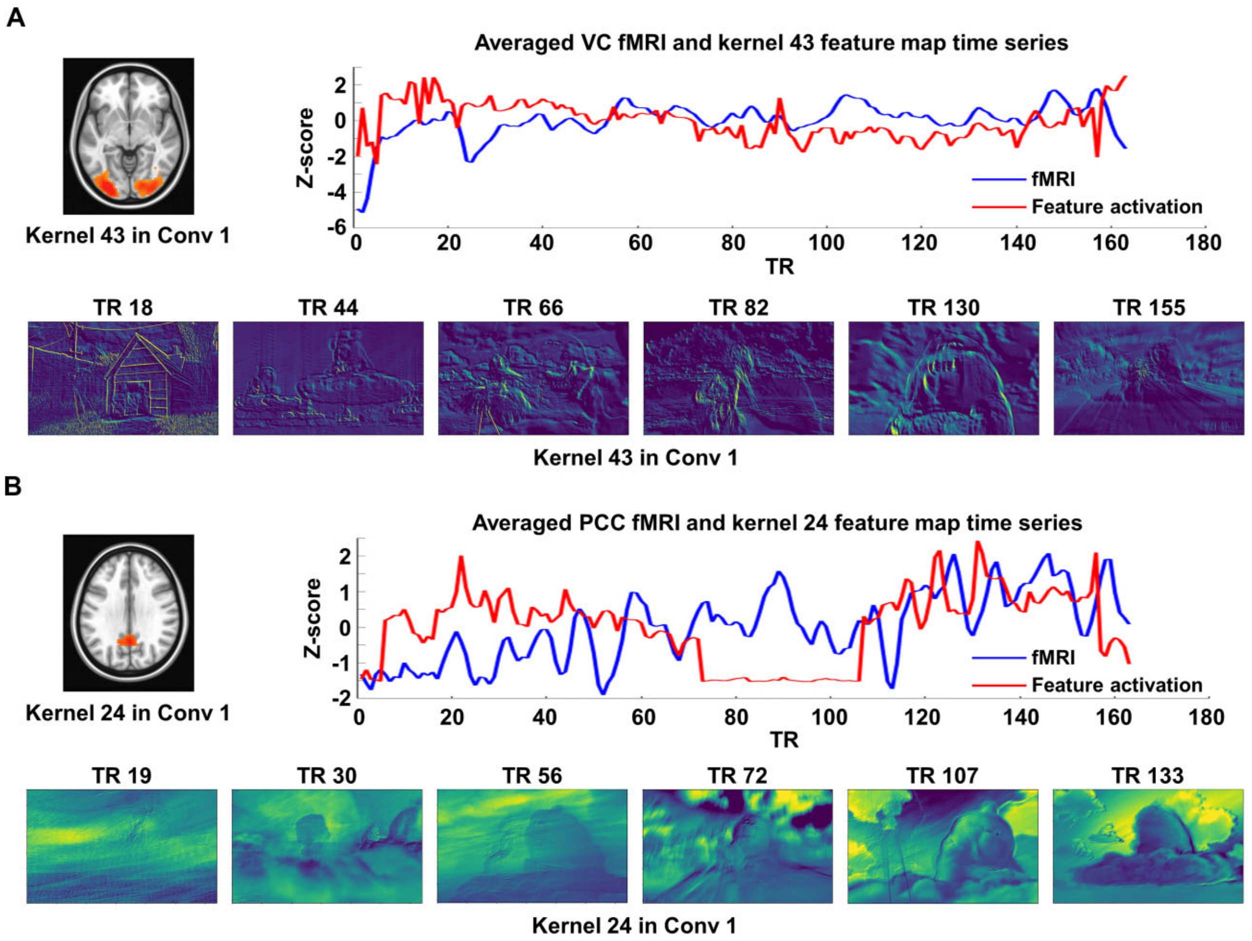

Figure 5: Reverse analysis of (**A**) the visual cortex and (**B**) posterior cingulate cortex. Activation maps were obtained as GLM results for (A) kernel 43 related to visual cortex and (B) kernel 24 related to posterior cingulate cortex. The z-scores of the averaged fMRI time series for each brain region are shown in blue, and the averaged feature map time series extracted from each kernel (kernel 43 or 24) are shown in red. The feature maps are for the TR times associated with the peaks of the averaged fMRI time series.

between specific brain regions and features, kernels that simultaneously activate multiple brain regions were excluded.

The visual cortex emerged as the most predictable region associated with features from the first convolutional layer. Specifically, we observed kernels 7, 43, and 63 of the first convolutional layer showing activation only in the visual cortex (Fig. 5A). Figure 5A showcases the BOLD time series in the visual cortex, allowing us to pinpoint the time points of peak regional activations. Furthermore, the feature activation maps corresponding to these peak time points for kernel 43 are also shown. These maps validate that the features associated with this particular kernel primarily emphasize edges or boundaries of objects, disregarding both the background and main characters.

The activations of the posterior cingulate cortex by features from the first convolutional layer were unexpected. This region is involved in the default mode network, which is typically associated with higher-order brain functions. Nonetheless, we successfully identified kernels 6, 19, 24, 42, 55, and 58 of the first convolutional layer that displayed activations only in the posterior cingulate cortex. Subsequently, we calculated the averaged BOLD time series within this region and identified the feature activation maps corresponding to time points exhibiting high BOLD activations (Fig. 5B). Intriguingly, these maps indicate that the features within these kernels primarily relate to image backgrounds rather than the main characters.

The supramarginal gyrus displayed the highest probability of activations among all regions for the thirteenth convolutional layer. Previous studies have associated this region with empathy for pain (Richardson *et al.*, 2018). Within the supramarginal gyrus, we identified kernels 42, 103, 168, 276, 428, and 510 of the thirteenth convolutional layer that exhibited activations. Subsequently, we extracted the time series of BOLD activations in this region (Fig. 6A). The feature activations in these kernels often appear blurry due to earlier convolutions and max-pooling layers. To enhance interpretability, we overlaid the feature activations with the input image. Observing the overlaid images, it becomes evident that the feature activations may involve multiple characters (at time point 42) or be linked to facial expressions (at time point 145). These findings suggest that the features within these kernels possess the capacity to represent higher-order social information.

The lateral occipital complex region related to overall object shape perception was also noticeably observed in the thirteenth convolutional layer. Unlike visual cortex, which is related to low-level visual elements, lateral occipital complex is associated with the task of comprehensively recognizing objects (Grill-Spector, 2001). Kernels 9, 43, and 128 in the thirteenth convolutional layer were related to lateral occipital complex. After extracting the averaged time series of BOLD activations, feature maps from the peaks of the time series were analyzed (Fig. 6B). Activation

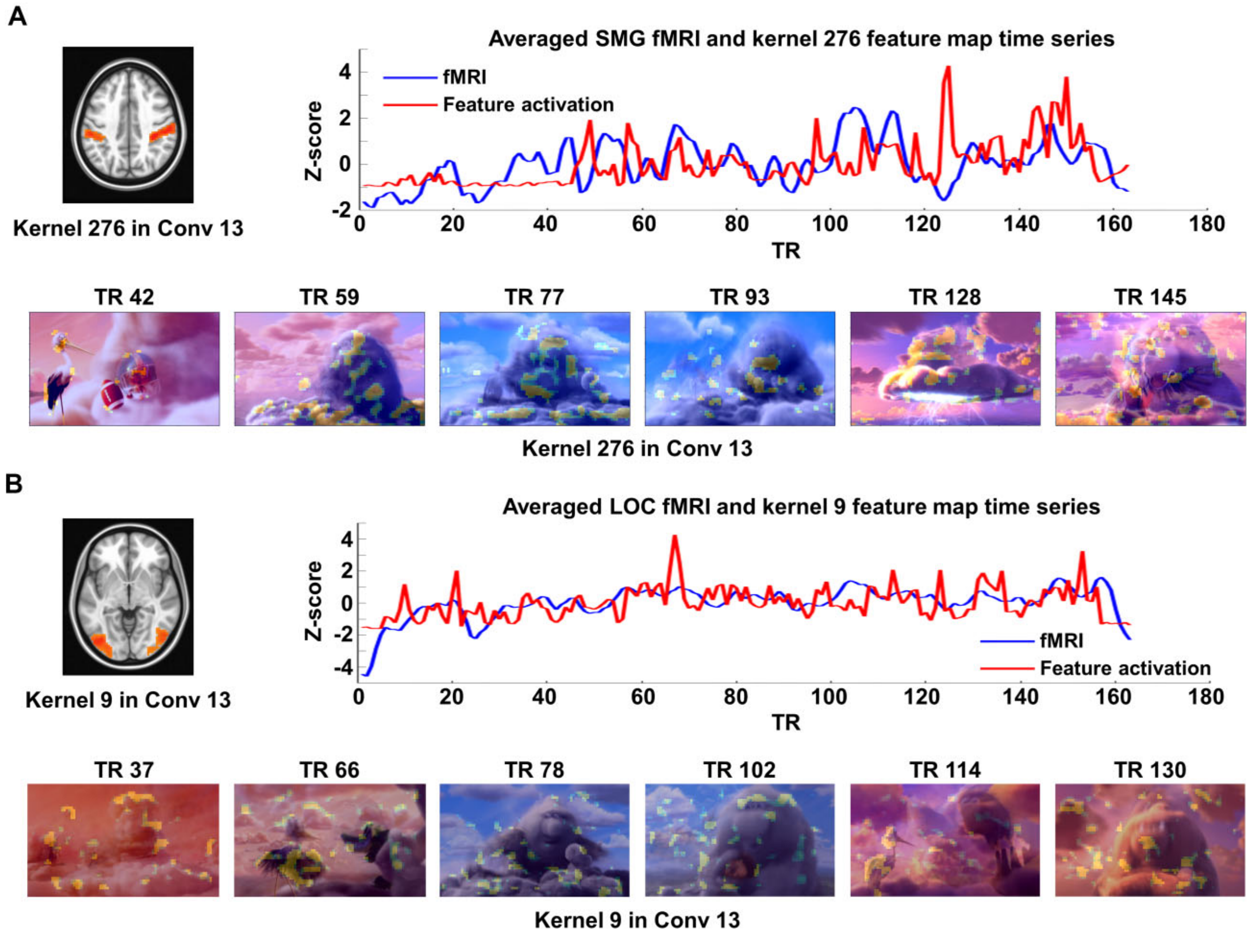

Figure 6: Reverse analysis of (**A**) supramarginal gyrus and (**B**) lateral occipital complex regions. Activation maps were obtained as GLM results for (A) kernel 276 for supramarginal gyrus and (B) kernel 9 for lateral occipital complex. The z-scores of the averaged fMRI time series for each brain region are shown in blue, and the averaged feature map time series extracted from each kernel (kernel 43 or 24) are shown in red. The feature maps are for the TR times related to the peaks of the averaged fMRI time series.

occurred across a variety of different objects (clouds, animals, and background objects).

Discussion

In this study, we employed a widely used CNN, VGG-16, to extract diverse visual features in a movie, and associated these feature activations with brain activations recorded through fMRI. The observed brain activations demonstrated to some extent a hierarchical pattern across convolutional layers. Lower convolutional layers showed stronger associations with posterior visual cortex and posterior cingulate cortex, whereas higher convolutional layers exhibited stronger associations with lateral occipital cortex and supramarginal gyrus. However, within a given convolutional layer the differentiation of different feature activations was limited, leading to many different feature activations associated with similar brain regions.

As a sanity check, we first examined the brain activations associated with the first layer of VGG-16, where the features are thought to reflect most basic local features of an image (Zeiler and Fergus, 2014; Krizhevsky *et al.*, 2017). As predicted, most of the associated brain activations were observed in the posterior occipital cortex, which corresponds to lower visual areas. The correspondence suggests, to some degree, similar visual features may be processed in shallow layers in the CNN and early visual brain areas. The current results also indicated limitations of using the temporal dynamic patterns to study brain activations. That is, the brain activation patterns from the different kernels of a layer were spatially highly similar and only showed a small number of distinct patterns. This makes sense given that the kernel activation time series from a layer were also highly correlated, and only a few principal components could explain most of the variance (Fig. 2A). Even though different kernels extract distinct features, the changes in activations across different input may be highly similar, e.g. edges in different directions may produce highly correlated activations over time. The temporal smoothing with hemodynamic response function may also contribute to the high correlations. Nevertheless, the factors that cause temporal similarity in the CNN resemble those measured with BOLD fMRI, which also suffer from poor specificity in measuring visual responses in brain. Alternative approaches, such as multivoxel pattern analysis (Kriegeskorte *et al.*, 2008) may be more effective to study the brain representation of different visual features.

Surprisingly, a small number of kernels in the first convolution layer were associated with brain activation in the posterior cingulate cortex, which is part of the default mode network (Raichle *et al.*, 2001). The default mode network is thought to be situated in a higher hierarchy of organization (Margulies *et al.*, 2016), and the associated functions are usually related to internalization and higher social functions (Buckner and DiNicola, 2019). However, recent studies also showed its involvements in naturalistic perception (Brandman *et al.*, 2021). The current results provide further insights into task related activations in the default mode. Specifically, the current results demonstrated that the activations of visual features from the first layer of VGG-16 can be related to the activations in the posterior cingulate cortex. Feature analysis further indicated that the visual activations of these kernels were associated with the background, rather than the main characters of a scene. The activations in the posterior cingulate cortex may be related to the processing of unattended features in the background, or may be negatively correlated with the level of attentions (Kaefer *et al.*, 2022). Nevertheless, the current results indicate that simple visual features may be linked to brain activity in the posterior cingulate cortex. Reverse inference of functions in the posterior cingulate cortex requires extra caution.

As the convolutional layer progresses deeper, the brain regions linked to kernel activations shift in a forward and upward direction within the brain. Specifically, there is a greater likelihood of association with the lateral occipital complex and the supramarginal network, both of which exhibit high inter-participant correlations (Di and Biswal, 2020). The lateral occipital complex plays a role in higher-order visual processing associated with object recognition (Grill-Spector *et al.*, 2001), distinguishing itself from lower visual areas specialized in low-level visual features like texture (Malach *et al.*, 1995). Hence, there seems to be a loose correspondence between the hierarchy of CNNs and the visual processing system in the brain. The reverse analysis further demonstrated that feature activations linked to the lateral occipital complex were typically situated around the main characters or objects in a scene (Fig. 6B). These findings align with a recent study demonstrating that responses from the lateral occipital complex were significantly predictive for scene, object, and action recognition in videos during movie viewing (McMahon *et al.*, 2023).

The supramarginal gyrus serves as a crucial brain region for the recognition and comprehension of others' emotions and pain (Lamm *et al.*, 2011; Silani *et al.*, 2013; Di and Biswal, 2020). The correlations between brain activations in the supramarginal gyrus and feature activations in deeper layers suggested that deeper layers of VGG-16 may be able to extract social information. In the reverse analysis, the feature activations by the relevant thirteenth layer kernels may at times involve two characters and at other times be associated with a single main character (Fig. 6A). Further investigations are needed to examine the activation characteristics of these thirteenth layer kernels across various images depicting social interactions, thereby validating our initial observations. Nonetheless, these results suggest that CNNs hold promise in representing high-order social information.

Several limitations in the current analysis should be taken into account. First, the utilization of an animated video clip as a naturalistic stimulus might introduce a potential bias. Given its artificial nature, CNNs may be more predisposed to its specific characteristics. The transferability of such models to videos featuring real human actors remains an open and challenging question. Previous studies have demonstrated differences in brain activations when participants watch animated versus real-life movies (Han *et al.*, 2005; Di *et al.*, 2022). Second, this study utilized a CNN to extract information from 2D frames of a video, revealing that higher layers of VGG-16 may capture higher-order social information. However, social interactions often entail dynamic changes over time, and models considering temporal dependencies, such as recurrent neural networks, may be more apt for extracting such information. Nonetheless, the inclusion of temporal dependencies increases model complexity. Future investigations could explore alternative model options to enhance our comprehension of social interactions. Finally, this study analyzed a small dataset with only adult participants. How brain activations vary across individuals, particularly in different age groups or as influenced by biological sex, remains largely unknown. These factors are crucial and warrant further investigation.

Conclusion

We have explored the utilization of a CNN model, specifically VGG-16, to extract diverse features at different levels. We linked these feature activations with brain activations measured through fMRI. Our analysis unveiled intricate relationships between brain activations and various layers of the CNN. Lower convolutional layers predominantly correlated with lower visual areas, while some exhibited associations with the posterior cingulate cortex, a component of the default mode network. By contrast, higher convolutional layers showed stronger associations with lateral occipital regions and the supramarginal gyrus, the latter being linked to higher-order social processing such as empathy for pain. Despite a high correlation in the temporal dynamics of kernel activations within the same layer, our findings suggest that kernels in deeper layers may signify more complex aspects of social interaction. The features extracted in this study provide a basis for future comparisons with alternative deep neural network models. Additionally, they hold promise for quantifying the content of movies and advancing our comprehension of the neural processes underlying cinematic experiences.

Author contributions

Wonbum Sohn (Conceptualization, Formal analysis, Investigation, Software, Visualization, Writing – original draft), Xin Di (Conceptualization, Funding acquisition, Project administration, Software, Writing – review & editing), Zhen Liang (Writing – review & editing), Zhiguo Zhang (Writing – review & editing), and Bharat B. Biswal (Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing)

Conflict of interest

One of the authors, Bharat B. Biswal, is also the associate editor of *Psychoradiology*. He was blinded from reviewing or making decisions on the manuscript.

Acknowledgements

This work was supported by (US) National Institute of Mental Health grants to Xin Di (R15MH125332) and Bharat B. Biswal (R01MH131335).

Data and code availability

The fMRI data used in this study is public data sets and is available on openneuro (https://openneuro.org/; accession #: ds000228). The codes in this study are available upon a reasonable request to the corresponding author.

References

Bartels A, Zeki S, Logothetis NK (2008) Natural vision reveals regional specialization to local motion and to contrast-invariant, global flow in the human brain. *Cereb Cortex* **18**:705–17.

Brandman T, Malach R, Simony E (2021) The surprising role of the default mode network in naturalistic perception. *Commun Biol* **4**: 1–9.

Buckner RL, DiNicola LM (2019) The brain's default network: updated anatomy, physiology and evolving insights. *Nat Rev Neurosci* **20**:593–608.

Çelik E, Keles U, Kiremitçi İ, *et al.* (2021) Cortical networks of dynamic scene category representation in the human brain. *Cortex* **143**:127–47.

Chen P-HA, Jolly E, Cheong JH, *et al.* (2020) Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *Neuroimage* **216**: 116851.

Deng J, Dong W, Socher R, *et al.* (2009) Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA: IEEE, 248–55.

Di X, Biswal BB (2015) Characterizations of resting-state modulatory interactions in the human brain. *J Neurophysiol* **114**: 2785–96.

Di X, Biswal BB (2020) Intersubject consistent dynamic connectivity during natural vision revealed by functional MRI. *Neuroimage* **216**:116698.

Di X, Zhang Z, Xu T, *et al.* (2022) Dynamic and stationary brain connectivity during movie watching as revealed by functional MRI. *Brain Struct Funct* **227**:2299–312,

Friston KJ, Williams S, Howard R, *et al.* (1996) Movement-related effects in fMRI time-series: movement artifacts in fMRI. *Magn Reson Med* **35**:346–55.

Grill-Spector K, Kourtzi Z, Kanwisher N (2001) The lateral occipital complex and its role in object recognition. *Vision Res* **41**: 1409–22.

Han S, Jiang Y, Humphreys GW, *et al.* (2005) Distinct neural substrates for the perception of real and virtual visual worlds. *Neuroimage* **24**:928–35.

Hasson U, Nir Y, Levy I, *et al.* (2004) Intersubject synchronization of cortical activity during natural vision. *Science* **303**:1634–40.

Hu W, Zhang Z, Zhao H, *et al.* (2023) EEG microstate correlates of emotion dynamics and stimulation content during video watching. *Cereb Cortex* **33**:523–42.

Jiahui G, Feilong M, Visconti di Oleggio Castello M, *et al.* (2022) Not so fast: limited validity of deep convolutional neural networks as in silico models for human naturalistic face processing. *J Vision* **22**:3714.

Kaefer K, Stella F, McNaughton BL, *et al.* (2022) Replay, the default mode network and the cascaded memory systems model. *Nat Rev Neurosci* **23**:628–40.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis—connecting the branches of systems neuroscience. *Front Syst Neuroscience* **2**:4.

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. *Commun ACM* **60**:84–90.

Lamm C, Decety J, Singer T (2011) Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage* **54**:2492–502.

Malach R, Reppas JB, Benson RR, *et al.* (1995) Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc Natl Acad Sci USA* **92**:8135–9.

Margulies DS, Ghosh SS, Goulas A, *et al.* (2016) Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc Natl Acad Sci USA* **113**:12574–9.

McMahon E, Bonner MF, Isik L (2023) Hierarchical organization of social action features along the lateral visual pathway. *Curr Biol* **33**:5035–5047.e8. e8.

Nastase SA Gazzola V Hasson U Keysers C (2019) Measuring shared responses across subjects using intersubject correlation.. *Soc Cogn Affect Neurosci* **14**:667–85.

Raichle ME, MacLeod AM, Snyder AZ, *et al.* (2001) A default mode of brain function. *Proc Natl Acad Sci USA* **98**:676–82.

Rao H, Wang J, Tang K, *et al.* (2007) Imaging brain activity during natural vision using CASL perfusion fMRI. *Hum Brain Mapp* **28**: 593–601.

Raz G, Winetraub Y, Jacob Y, *et al.* (2012) Portraying emotions at their unfolding: a multilayered approach for probing dynamics of neural networks. *Neuroimage* **60**:1448–61.

Richardson H, Lisandrelli G, Riobueno-Naylor A, *et al.* (2018) Development of the social brain from age three to twelve years. *Nat Commun* **9**:1027.

Silani G, Lamm C, Ruff CC, *et al.* (2013) Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *J Neurosci* **33**:15466–76.

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society*, 1–14.

Sun Y, Ma J, Huang M, *et al.* (2022) Functional connectivity dynamics as a function of the fluctuation of tension during film watching. *Brain Imag Behav* **16**:1260–74.

Zeiler MD, Fergus R. (2014) Visualizing and understanding convolutional networks. In: D Fleet, T Pajdla, B Schiele, T Tuytelaars (eds). *Computer Vision—ECCV 2014. Lecture Notes in Computer Science*. Chamonix: Springer International Publishing. p. 818–33.

Received: 19 July 2024; **Revised:** 9 September 2024; **Accepted:** 31 October 2024