Multimodal deep neural decoding of visual object representation in humans

Noriya Watanabe[1], Kosuke Miyoshi[2], Koji Jimura[3], Daisuke Shimane[1], Ruedeerat Keerativittayayut[1,4], Kiyoshi Nakahara[1], Masaki Takeda[1*]

[1]Research Center for Brain Communication, Kochi University of Technology, Kami, Kochi, 782-8502, Japan
[2]Narrative Nights, Inc., Yokohama, Kanagawa, 236-0011, Japan
[3]Department of Biosciences and Informatics, Keio University, Yokohama, Kanagawa, 223-0061, Japan
[4]Present address: Chulabhorn Royal Academy, Bangkok, 10210, Thailand

[*]Correspondence should be addressed to:
Masaki Takeda, Ph.D.
Research Center for Brain Communication,
Kochi University of Technology,
185, Miyanokuchi, Tosayamada-cho, Kami-shi
Kochi, 785-8502, Japan
Email: takeda.masaki@kochi-tech.ac.jp
Phone: +81-887-57-2237

Keywords: fMRI, EEG, convolutional neural network, visual system

Abstract

Perception and categorization of objects in a visual scene are essential to grasp the surrounding situation. However, it is unclear how neural activities in spatially-distributed brain regions, especially in the aspect of temporal dynamics, represent visual objects. To address this issue, we explored spatial and temporal organization of visual object representations using concurrent functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), combined with neural decoding by deep neural networks (DNNs). Visualization of the fMRI DNN revealed that visual categorization (faces or non-face objects) occurred in brain-wide cortical regions, including the ventral temporal cortex. Interestingly, the EEG DNN valued the earlier phase of neural responses for categorization and the later phase of neural responses for sub-categorization. Combination of both DNNs improved classification performance for both categorization and sub-categorization. These deep learning-based results demonstrate a categorization principle in which visual objects are represented in a spatially organized and coarse-to-fine manner.

Introduction

Visual perception is a rapid cognitive process that allows immediate judgement of safety/risk, enabling to quick planning and acting. The recognition and categorization of the visual scene requires a series of processing stages in the ventral visual processing stream (Grill-Spector and Weiner, 2014; Grill-Spector et al., 2017; Hung et al., 2005; Logothetis and Sheinberg, 1996; Peelen and Downing, 2017). The last stage of the ventral stream, where both perceptual and mnemonic signals meet, plays a crucial role in this cognitive process by referring the perceived object to the neural representation of memorized objects stored in the long-term memory (Landi et al., 2021; Suzuki and Naya, 2014; Takeda et al., 2018).

The functional role of the ventral stream on visual categorization has been investigated by neuropsychological studies. Lesions to the ventral stream can cause various forms of agnosia depending on the location of the core lesion (Barton, 2011; Cavina-Pratesi et al., 2010; Gainotti and Marra, 2011; Konen et al., 2011; Schiltz et al., 2006), suggesting module-like spatial organization for visual categories, such as faces and objects. These observations are further supported by accumulating evidence from functional magnetic resonance imaging (fMRI) in humans, such as the identification of category-selective regions for faces (Haxby et al., 2001; Ishai et al., 1999; Kanwisher et al., 1997), objects (Haxby *et al.*, 2001; Ishai *et al.*, 1999; Kourtzi and Kanwisher, 2000; Malach et al., 1995), and places (Epstein and Kanwisher, 1998) in not only the ventral stream, but also other cortical regions. Spatial organization in the categorization was also demonstrated in non-human primates (Kriegeskorte et al., 2008; Tsao et al., 2003; Tsao et al., 2006; Turesson et al., 2012).

In contrast, neural representations for categorical subdivision (sub-categorical representation), especially in the temporal aspects, remain unclear. At the finest spatial scale, single neurons show selective responses to identities within a single category in non-human primates (Miyashita and Chang, 1988; Tanaka, 1996). At the larger spatial scale, high-frequency local field potential carried spike-coupled information, such as the species, view, and identity of the face (Miyakawa et al., 2018). In human studies, extensive evidence has been obtained from magneto-/electroencephalography (MEG/EEG) (Carlson et al., 2013; Cichy et al., 2014; Dobs et al., 2019; Liu et al., 2002; Zheng et al., 2012) (see also fMRI studies for sub-categorical representation (Kriegeskorte et al., 2007; Nestor et al., 2011)). However, the temporal trajectory of categorical and sub-categorical representation, that is, whether our brains process visual object categorization in a fine-to-coarse or coarse-to-fine manner, is still controversial

(Carlson *et al.*, 2013; Cichy *et al.*, 2014; Dobs *et al.*, 2019; Liu *et al.*, 2002; Mack and Palmeri, 2011).

Unraveling the whole picture of the spatiotemporal profile of categorical and sub-categorical representations of visual objects is indispensable for understanding the computational model of visual object processing in the brain. To address this issue, we introduce two technical implementations here. First, we employed concurrent fMRI and EEG recordings in humans, to obtain neural responses to visual stimuli in high spatiotemporal resolution. Second, we utilized the deep neural network (DNN) to resolve spatiotemporal patterns of the neural activity responsible for the categorical/sub-categorical classification of visual objects. We hypothesized that (1) fMRI data involve spatially resolved patterns of neural activity that contribute to visual object classification between categories (faces or objects), and (2) EEG data involve temporally resolved patterns of neural activity that contribute to visual object classification not only between categories but also within a category (sub-categorization; male faces or female faces, and natural objects or artificial objects), both of which would be processed at different periods (Figure 1A). These hypotheses were tested using newly-developed DNN classifiers for multimodal neural data recorded by concurrent fMRI and EEG (Figure 2) as well as the univariate analysis for fMRI and EEG data.

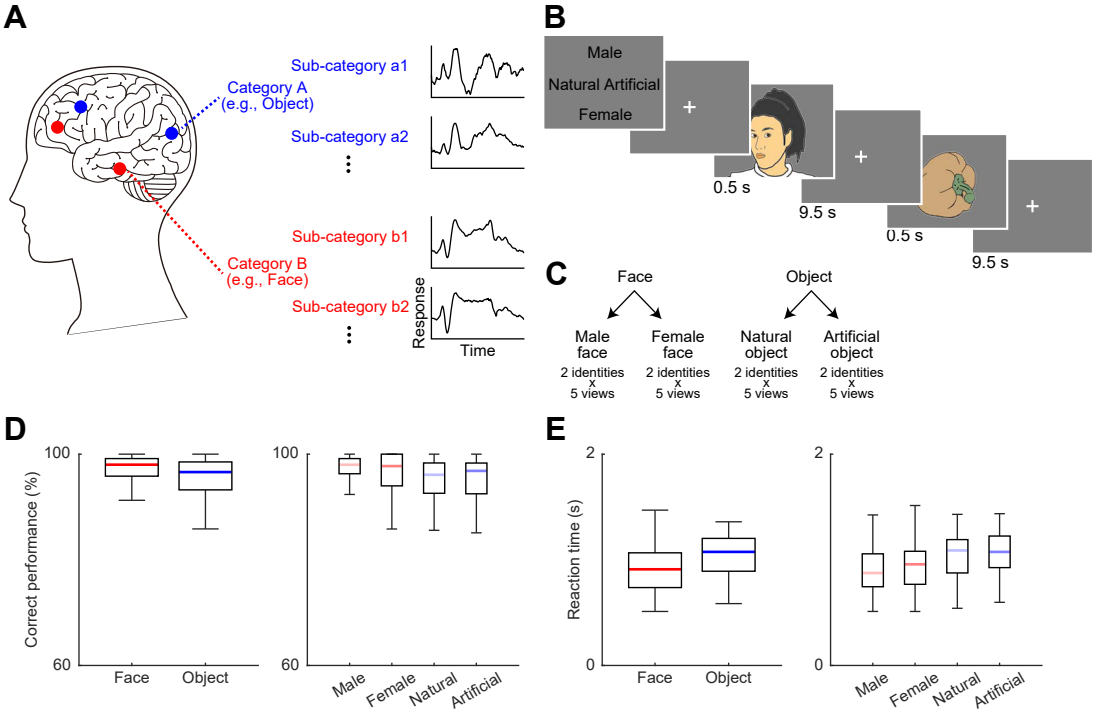
A
Sub-category a1
Category A
(e.g., Object)
Sub-category a2
Sub-category b1
Category B
(e.g., Face)
Sub-category b2
Response
Time
B
Male
Natural Artificial
Female
0.5 s
9.5 s
0.5 s
9.5 s
C
Face
Object
Male face
Female face
Natural object
Artificial object
2 identities x 5 views
2 identities x 5 views
2 identities x 5 views
2 identities x 5 views
D
Correct performance (%)
100
60
Face
Object
Male
Female
Natural
Artificial
E
Reaction time (s)
2
0
Face
Object
Male
Female
Natural
Artificial

Figure 1

Figure 1. Experimental paradigm and behavioral performance. (A) A hypothesis for the neural representation of visual object categorization and sub-categorization. Categorical information, such as identification of faces or objects, is mainly represented in the spatial domain, whereas sub-categorical information such as male faces or female faces is mainly represented in the temporal domain. **(B)** Sequence of a task run. At the start of a task run, the association between visual objects and buttons is instructed. Each visual stimulus from four sub-categories (male face, female face, natural object, artificial object stimuli) is presented for 500 ms, with a 10-s interval. **(C)** Exemplar visual stimuli. Face and object images are selected from ATR Facial Expression Database DB99 and ALOI (Amsterdam Library of Object Images), respectively. See also Figure 1 – figure supplement 1 for their low-level visual features. **(D–E)** Correct performance (D) and response latency (E) for each category and sub-category. **(D)** Differences in the correct performance in both face sub-category and object sub-category are not statistically significant (right, Wilcoxon signed rank test, $z < 0.4$, $P > 0.68$), though performance in the face trial is higher than that in the object trial (left, face and object; $z = 4.1$, $P = 4.19\times10^{-5}$). **(E)** Left, reaction times to face stimuli are significantly faster than those to object stimuli (Wilcoxon signed rank test, $z = 5.56$, $P = 2.63\times10^{-8}$). Right, within face sub-category, reaction times to male face are faster than those to female face ($z = 3.50$, $P = 4.63\times10^{-4}$), while difference in reaction times within object sub-category is not significant ($z = 0.71$, $P = 0.48$). $N = 50$ participants. Boxplots show median (thick line), $25^{th}/75^{th}$ percentile (box), and minimum/maximum (whisker).

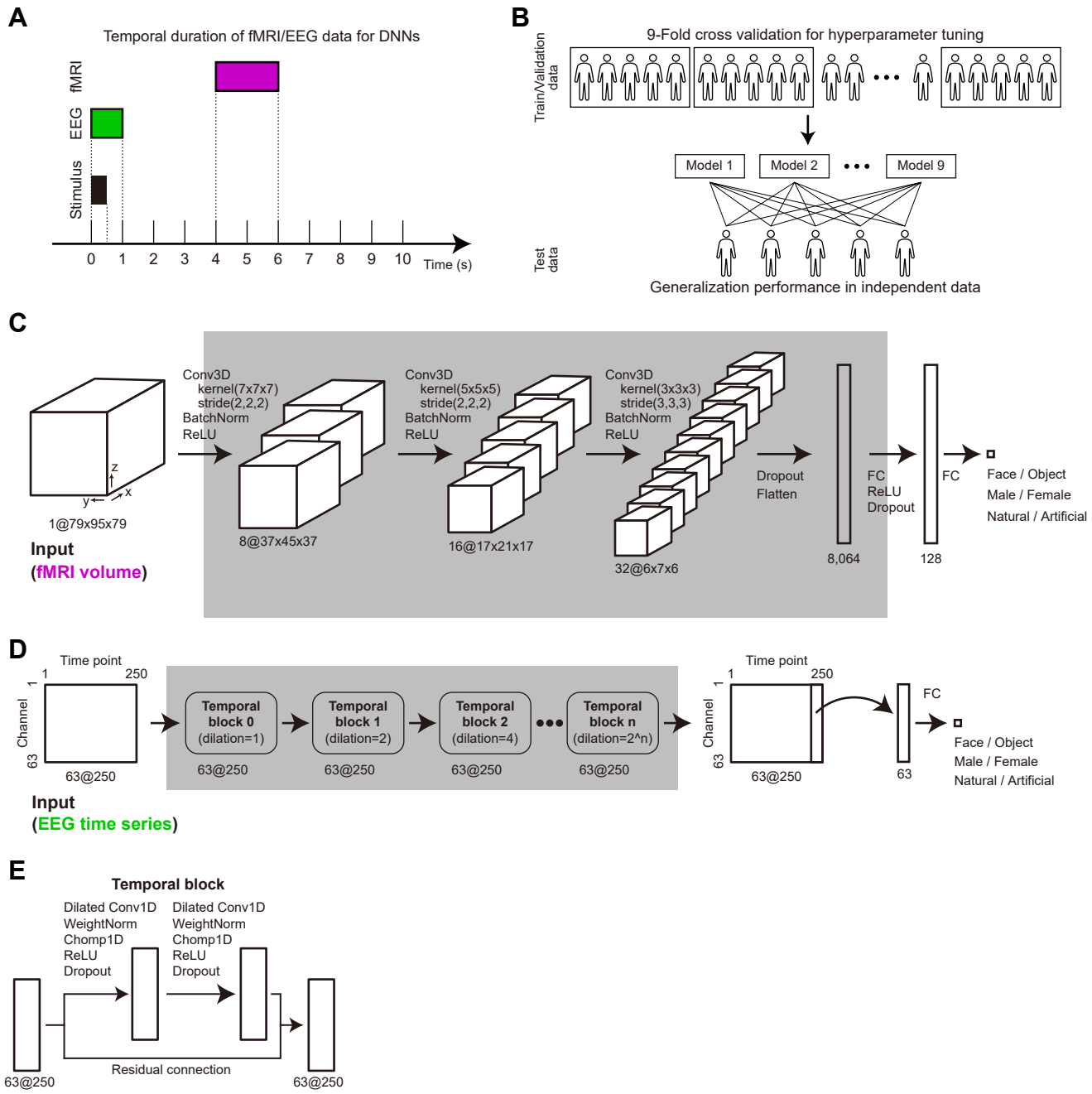
A
Temporal duration of fMRI/EEG data for DNNs
fMRI
EEG
Stimulus
0 1 2 3 4 5 6 7 8 9 10
Time (s)
B
9-Fold cross validation for hyperparameter tuning
Train/Validation data
Model 1
Model 2
Model 9
Test data
Generalization performance in independent data
C
Conv3D
kernel(7x7x7)
stride(2,2,2)
BatchNorm
ReLU
Conv3D
kernel(5x5x5)
stride(2,2,2)
BatchNorm
ReLU
Conv3D
kernel(3x3x3)
stride(3,3,3)
BatchNorm
ReLU
1@79x95x79
Input
(fMRI volume)
8@37x45x37
16@17x21x17
32@6x7x6
Dropout
Flatten
8,064
FC
ReLU
Dropout
128
FC
Face / Object
Male / Female
Natural / Artificial
D
Time point
Channel
63@250
Temporal block 0 (dilation=1)
Temporal block 1 (dilation=2)
Temporal block 2 (dilation=4)
Temporal block n (dilation=2^n)
Input
(EEG time series)
FC
63
Face / Object
Male / Female
Natural / Artificial
E
Temporal block
Dilated Conv1D
WeightNorm
Chomp1D
ReLU
Dropout
Residual connection
63@250

Figure 2

Figure 2. DNN models for multimodal neural data. (A) Temporal relationship across stimulus presentation and fMRI/EEG data windows used for DNN. fMRI data were taken within 4–6 s from stimulus onset to consider the delayed hemodynamic response of the BOLD signal. EEG data were taken during the period from stimulus onset to 0.5 s after stimulus offset. **(B)** Schematics of the train/validation/test architecture. In the training/validation dataset (*N* = 45 participants, 1-fold data include data from *N* = 5 participants), we applied 9-fold cross-validation to tune hyperparameters and constructed learned DNN models. Then, the generalization performances of these models for the independent test dataset were assessed (*N* = 5 participants). **(C)** A DNN model for fMRI data. The model includes three convolution layers and two fully connected layers. **(D)** A DNN model for EEG data. The model includes four temporal blocks and one fully connected layer. **(E)** A temporal block of EEG DNN. This block includes two dilated convolution layers with a residual connection.

Results

Behavioral results

Participants performed a visual object classification task, in which they judged whether the presented visual stimuli belonged to either the male face, female face, natural object, or artificial object image category (Figure 1B–C). To minimize the effect of low-level visual features of stimuli on differential brain activations and resultant differences in decoding performances, we controlled for contrast and luminance of images (Stigliani et al., 2015) (Figure 1 – figure supplement 1; see Materials and Methods for details).

Participants almost perfectly categorized the presented visual stimuli (Figure 1D; $N = 50$). Performance was not significantly different between male face and female face trials and between natural object and artificial object trials (Wilcoxon signed rank test, $z < 0.4$, $P > 0.68$). Thought reaction times to male face were faster than those to female face ($z = 3.50$, $P = 4.63\times10^{-4}$), there was no significant difference in reaction times within the object sub-category ($z = 0.71$, $P = 0.48$) (Figure 1E).

Spatial organization of category-specific information revealed by fMRI activations

We explored brain regions associated with categorical processing based on standard univariate general linear model (GLM) analysis of fMRI data (Figure 3, Figure 3 – figure supplement 1). Robust activations for faces compared with those for objects, were observed in various regions, including the temporal occipital fusiform cortex (TOF), which includes the region known as the fusiform face area (FFA), and lateral occipital complex (LOC) in the ventral temporal cortex (VTC), as well as the occipital pole, angular gyrus, frontal medial cortex, amygdala, frontal orbital cortex, precuneous (PcC), and middle temporal gyrus (Figure 3A). In contrast, the brain regions activated by objects included the TOF, LOC, middle frontal gyrus, insula, superior frontal gyrus, thalamus, and paracingulate gyrus. These results were consistent with previous studies (Haxby *et al.*, 2001; Ishai *et al.*, 1999; Kanwisher *et al.*, 1997; Kourtzi and Kanwisher, 2000; Malach *et al.*, 1995).

Subsequently, we explored brain regions associated with sub-categorical processing (Figure 3 – figure supplement 2, 3). The face gender contrast (male face vs female face) revealed weak activations; there were no male face dominant clusters and three female face dominant clusters (Figure 3B). The contrast between natural and artificial objects also showed weak activations; there were no natural object dominant clusters and two artificial object dominant clusters (Figure 3C). These results suggest that the difference in sub-categorical processing of visual objects is hardly detected at the spatiotemporal resolution of fMRI used in this study.

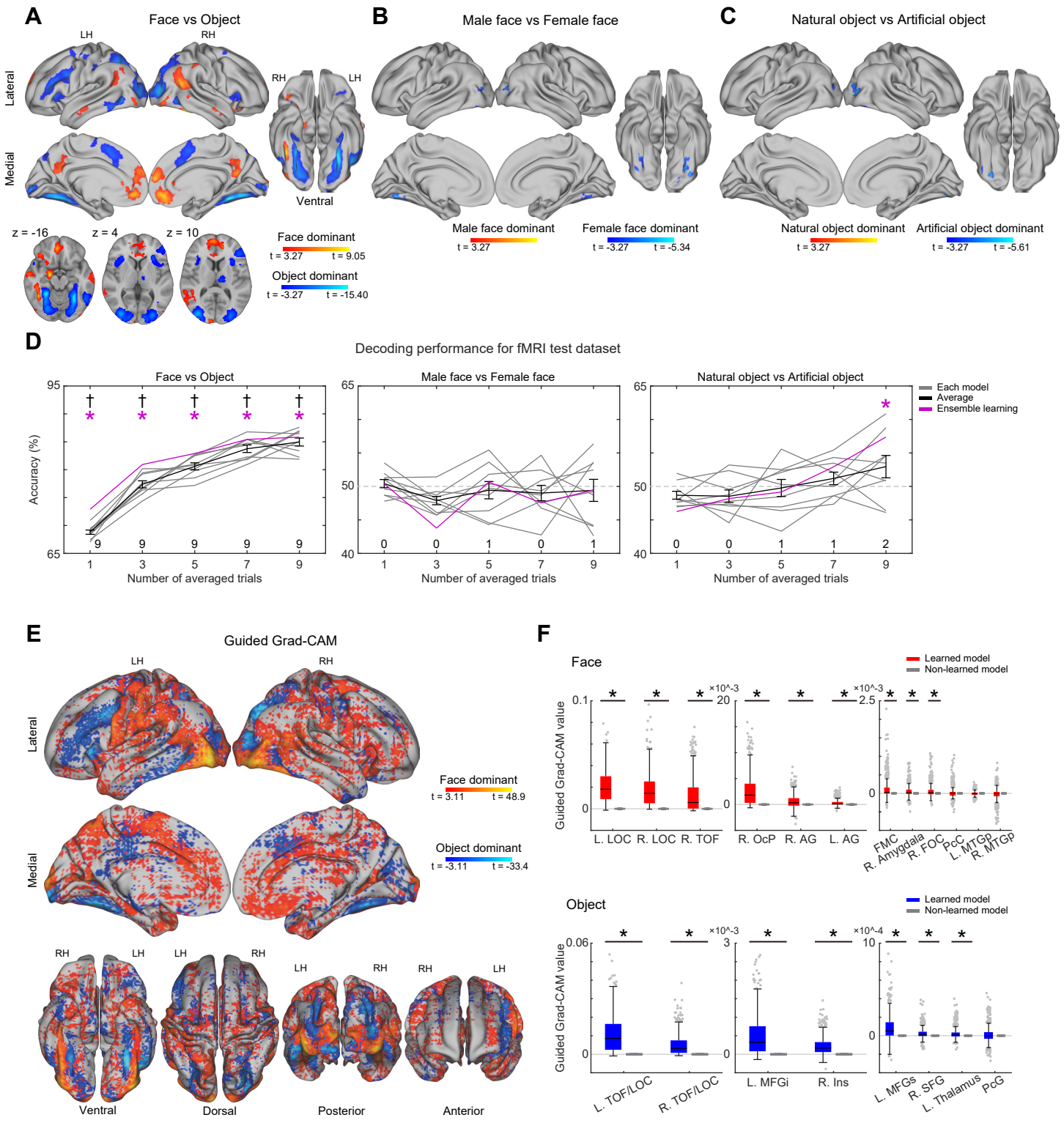
A
Face vs Object
LH
RH
Lateral
Medial
Ventral
z = -16
z = 4
z = 10
Face dominant
t = 3.27
t = 9.05
Object dominant
t = -3.27
t = -15.40
B
Male face vs Female face
Male face dominant
t = 3.27
Female face dominant
t = -3.27
t = -5.34
C
Natural object vs Artificial object
Natural object dominant
t = 3.27
Artificial object dominant
t = -3.27
t = -5.61
D
Decoding performance for fMRI test dataset
Face vs Object
Male face vs Female face
Natural object vs Artificial object
Accuracy (%)
Number of averaged trials
Each model
Average
Ensemble learning
E
Guided Grad-CAM
Face dominant
t = 3.11
t = 48.9
Object dominant
t = -3.11
t = -33.4
Ventral
Dorsal
Posterior
Anterior
F
Face
Learned model
Non-learned model
Guided Grad-CAM value
L. LOC
R. LOC
R. TOF
R. OcP
R. AG
L. AG
FMC
R. Amygdala
R. FOC
PcC
L. MTGp
R. MTGp
Object
L. TOF/LOC
R. TOF/LOC
L. MFGi
R. Ins
L. MFGs
R. SFG
L. Thalamus
PcG

Figure 3

Figure 3. Spatial representation of categorization/sub-categorization for visual objects revealed by fMRI. (A–C) Activation map for the contrast between face and object (A), between male face and female face trials (B) and between natural object and artificial object trials (C). Color bars denote t-value. Significant brain regions are denoted using a threshold of $P < 0.05$ corrected with cluster-wise family-wise error rate based on a non-parametric permutation test. Identified clusters in the subcortical structures and the insula are shown in the horizontal sections. See also Figure 3 – figure supplement 1–3. **(D)** Generalization decoding performance for the test dataset. Ensemble accuracy (purple), as well as average accuracy ± S.D (thick line) across models, are shown as a function of the number of averaged trials. Thin lines represent accuracy in each model. *, binomial test for ensemble accuracy against chance level (50%), FDR corrected $q < 0.05$. †, t-test for accuracy distribution against chance level, FDR corrected $q < 0.05$. The number of significant models (binomial test, $q < 0.05$) is shown above the abscissa. **(E)** Spatial organization of the guided Grad-CAM values for the contrast of category (between face and object trials). The threshold of significance is set at $P < 0.05$ corrected with cluster-wise FWE based on a non-parametric permutation test. **(F)** ROI analysis. Colored and grayed bars denote the guided Grad-CAM values in learned and non-learned models, respectively. For display purpose, the results are plotted in separate panels according to the magnitude of values. *, t-test against values obtained from the non-learned model, FDR corrected $q < 0.001$. Values in non-learned models are nearly zeros. Boxplots show median (black line), $25^{th}/75^{th}$ percentile (box), nonoutlier minimum/maximum (whisker), and outlier (dot, data points outside 1.5× interquartile range). Cluster information is listed in Figure 3 – figure supplement 1. L, left; R, right; LH, left hemisphere; RH, right hemisphere; AG, angular gyrus; FMC, frontal medial cortex; FOC, frontal orbital cortex; Ins, insular cortex; LOC, lateral occipital cortex; MFG, middle frontal gyrus; MTG, middle temporal gyrus; OcP, occipital pole; PcC, precuneous cortex; PcG, paracingulate gyrus; SFG, superior frontal gyrus; TOF, temporal occipital fusiform cortex; i, inferior; s, superior; p, posterior.

Deep neural decoding of visual object categorization by fMRI DNN

To decode information about the visual object category and identify the spatial organization of responsible brain regions, we used a DNN classifier for fMRI data, which was based on three-dimensional convolution layers (Figure 2C). We trained the fMRI DNN using an fMRI volume acquired 4 s after the presentation of visual stimuli in each trial (Figure 2A). Training and validation of the DNN were conducted by 9-fold cross-validation, and the generalization performance of the resultant nine models was evaluated using an independent test dataset (Figure 2B; see Materials and Methods for details).

The ensemble learning method using a majority voting scheme (Rokach, 2009) revealed that classification performance in categorization reached statistical significance when trials were not averaged (72.90%; $q = 2.61\times10^{-72}$, binomial test against chance level 50%, corrected for multiple comparisons by false discovery rate (FDR) of $q < 0.05$) (Figure 3D). The performance improved as a function of the number of averaged trials, reaching 85.83% with an average of nine trials ($q = 3.89\times10^{-185}$). In contrast, both face sub-categorical and object sub-categorical classification were not significant in the averaged trial condition ($q > 0.21$), except for object sub-categorization when nine trials were averaged.

Two additional analyses were performed to examine the robustness of these results. In the first analysis, a t-test showed that for categorical classification, the performance distribution statistically reached above chance when trials were not averaged [68.80 ± 0.39% (mean ± s.e.m.); $t = 48.80$, $P = 3.44\times10^{-11}$, corrected by FDR $q < 0.05$]. The performance improved as a function of the number of averaged trials (linear regression, $R^2 = 0.82$, slope coefficient, $t = 13.99$, $P = 1.32\times10^{-17}$), reaching 84.96 ± 0.71% with an average of nine trials ($t = 49.04$, $P = 3.31\times10^{-11}$). The highest accuracy reached was 87.60% in the best performing DNN model. In contrast, both face sub-categorical and object sub-categorical classification did not reach statistical significance in either averaged trial conditions ($q > 0.26$).

In the second analysis, we counted the number of models showing significant performance against chance level using a binomial test (threshold at $P = 0.05$, FDR corrected). Although all nine models were significant in categorical classification across all averaged trial conditions, there were less than two models with significant sub-categorical classification. The results in the two additional analyses support the result of the ensemble learning method.

With respect to important spatial neural activity for visual object classification, guided Gradient-weighted Class Activation Mapping (Selvaraju et al., 2019) (guided Grad-CAM) showed that the fMRI DNN model valued brain-wide regions, including the frontal, parietal, temporal, and occipital cortices as well as subcortical structures in both face and object trials (Figure 3E; Figure 3 – figure supplement 4). Whole-brain activations identified in the univariate analysis and guided Grad-CAM were significantly correlated at the voxel-by-voxel level, although majority of voxels showed guided Grad-CAM values around zero, leading to a relatively low effect size (Figure 3 – figure supplement 5A; Spearman's rank correlation, $\rho = 0.048$, $t = 19.73$, $P < 0.001$ for face trial; $\rho = -0.21$, $t = -90.13$, $P < 0.001$ for object trial). The correlation coefficient increased when target voxels were masked by the maps of meta-analysis obtained from Neurosynth (Yarkoni et al., 2011) using the terms "face" and "object" (Figure 3 – figure supplement 5B; $\rho = 0.27$, $t = 46.27$, $P < 0.001$ for face trial; $\rho = -0.30$, $t = -56.21$, $P < 0.001$ for object trial). This suggests that voxels with nearly-zero values in the guided Grad-CAM tended to be located outside significant clusters in the univariate analysis. Indeed, the region-of-interest (ROI) analysis revealed that most of the identified clusters in the univariate analysis showed significant guided Grad-CAM values in both face and object trials, when comparing with those obtained in non-learned DNN models where weight values were assigned randomly (Figure 3F; Wilcoxon signed rank test against guided Grad-CAM values in non-learned models, FDR corrected $q < 0.05$). Similar results were obtained when ROIs were defined based on the anatomical structure or functional meta-analysis by Neurosynth (Figure 3 – figure supplement 6); among the three ROI-based analyses, consistent results were obtained in the VTC (LOC, TOF), FMC, and PcC for the face trial, and the VTC (LOC), and thalamus for the object trial.

These results indicate that, for the categorization between faces and objects, our DNN model paid attention to the neural signals in the distributed brain regions identified in the univariate analysis.

Temporal trajectory of categorical/sub-categorical information by EEG responses

To examine temporal characteristics of categorical and sub-categorical processing of visual objects, we analyzed the event-related potential (ERP) and power spectrogram of the EEG signals (Figure 4; Figure 4 – figure supplement 1). Regarding ERPs for categorization, face stimuli induced an early negative component (< 200 ms from stimulus onset) with a prolonged component (around 200–700 ms), wherein the polarity of the potential and its dynamics differed across channels. Meanwhile, object stimuli did not evoke the clear early negative component (Figure 4 – figure supplement 1A).

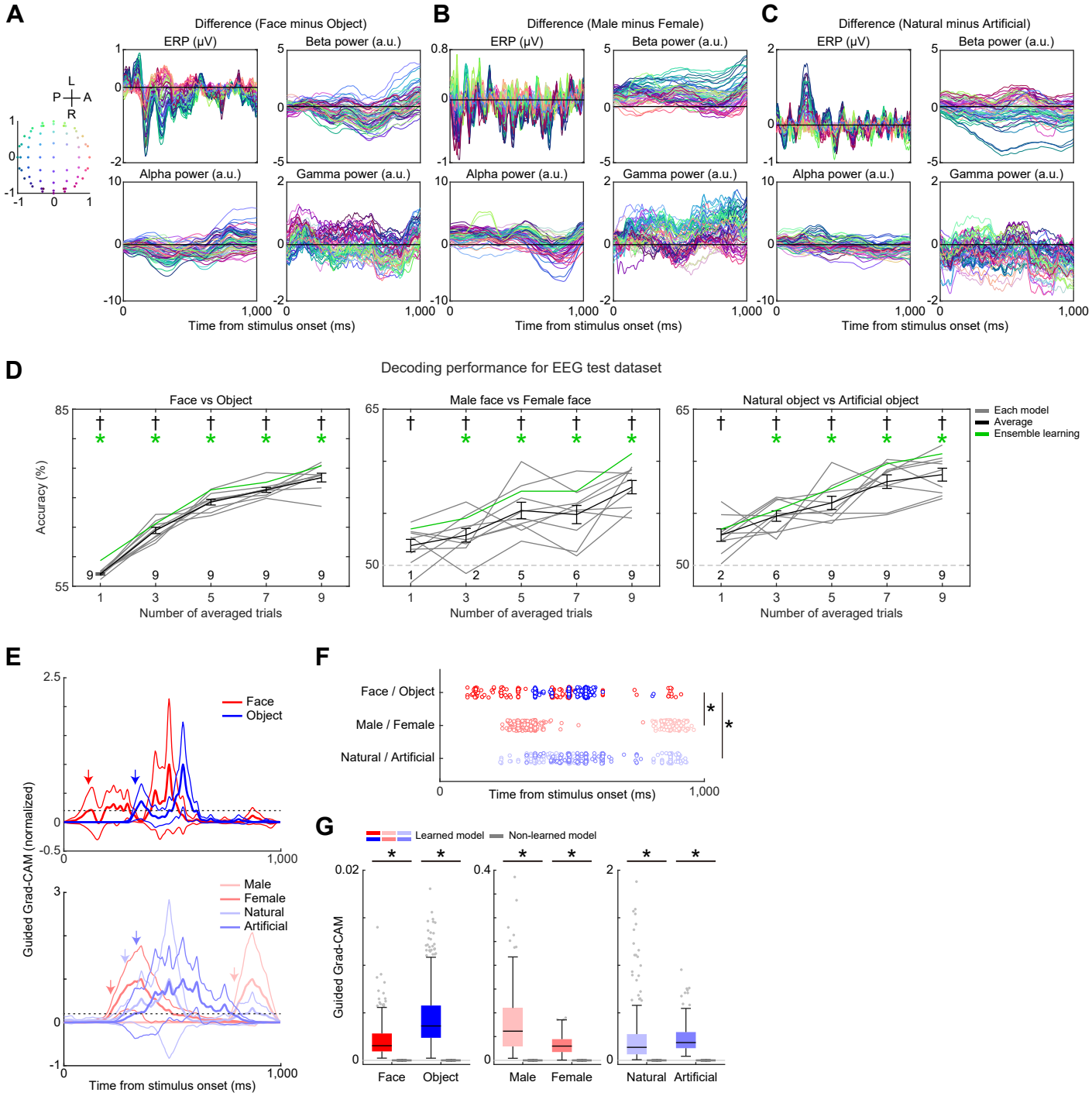
A
Difference (Face minus Object)
ERP (μV)
Beta power (a.u.)
Alpha power (a.u.)
Gamma power (a.u.)
Time from stimulus onset (ms)
B
Difference (Male minus Female)
C
Difference (Natural minus Artificial)
D
Decoding performance for EEG test dataset
Face vs Object
Male face vs Female face
Natural object vs Artificial object
Accuracy (%)
Number of averaged trials
Each model
Average
Ensemble learning
E
Face
Object
Male
Female
Natural
Artificial
Guided Grad-CAM (normalized)
Time from stimulus onset (ms)
F
Face / Object
Male / Female
Natural / Artificial
Time from stimulus onset (ms)
G
Learned model
Non-learned model
Guided Grad-CAM
Face
Object
Male
Female
Natural
Artificial

Figure 4

Figure 4. Temporal representation of categorization/sub-categorization for visual objects revealed by EEG. (A) Differential ERP and power spectrogram in the contrast of face minus object trials. Line colors correspond to the channel location depicted on the left side. A, anterior, P, posterior, R, right, L, left. **(B)** Differential ERP and power spectrogram in the contrast of male face minus female face trials. **(C)** Differential ERP and power spectrogram in the contrast of natural object minus artificial object trials. **(D)** Generalization decoding performance for the test dataset. Configurations are the same as in Figure 3d. **(E)** Temporal trajectories of the guided Grad-CAM values for each trial type. Data from all channels are averaged, and values are normalized to the maximum guided Grad-CAM value (=1). Dashed lines represent a threshold (20% of maximum value). Arrows depict maximum latency that first exceeded the threshold. **(F)** Temporal distribution of maximum latency for each trial (dot). Trials for face, object, male face, female face, natural object, and artificial object are separately colored. *, Tukey–Kramer test, $P < 0.05$ after Kruskal–Wallis test, $\chi^2 = 9.70$, $P = 0.0078$. **(G)** Comparison of the guided Grad-CAM values at the maximum latency between learned and non-learned models. *, t-test against values obtained from the non-learned model, FDR corrected $q < 0.001$. Note values in non-learned models are nearly zeros. Boxplots show median (black line), $25^{th}/75^{th}$ percentile (box), nonoutlier minimum/maximum (whisker), and outlier (dot, data points outside 1.5× interquartile range).

This led to differential ERPs between face and object trials (Figure 4A). The ERPs between faces and objects were significantly different at around 200–600 ms from stimulus onset at brain-wide (50/63) channels (Figure 4 – figure supplement 1B; cluster-defining threshold of $P < 0.05$, corrected significance level of $P < 0.05$). In contrast, the power spectrogram did not show clear differences (Figure 4A; Figure 4 – figure supplement 1B).

In contrast to categorization, spectral power was more informative in sub-categorization. Specifically, in face sub-category processing, male face stimuli elicited larger beta power than did female face stimuli at a prolonged time window (–1,000 ms). Meanwhile, these two stimuli did not evoke distinctive ERP responses (Figure 4B and Figure 4 – figure supplement 1C). In object sub-category processing, artificial stimuli evoked larger beta power than did natural stimuli at a prolonged time window (–1,000 ms) in several channels at the posterior part of the brain. Distinctive ERP responses were also observed at 200–300 ms (21/63 channels; Figure 4D and Figure 4 – figure supplement 1D).

Deep neural decoding of visual object categorization/sub-categorization by EEG DNN

To decode information about the visual object category and sub-category as well as to identify its temporal dynamics, we used a DNN classifier for EEG data, which was based on a temporal convolutional network (TCN) (Bai et al., 2018) (Figure 2D–E). We trained the EEG DNN using 63-ch EEG time series during 1 s after stimulus onset (Figure 2A). The ensemble learning method revealed that classification performance in categorization statistically reached above chance when trials were not averaged (59.33%; binomial test against chance level 50%, $q = 5.75\times10^{-13}$, corrected for multiple comparisons by FDR $q < 0.05$) (Figure 4D). The performance improved as a function of the number of averaged trials, reaching 75.41% with an average of nine trials ($q = 4.95\times10^{-89}$). Regarding face and object sub-categorization, classification performance improved from 53.50% (no trial average; $q = 0.056$) to 60.75% (average of 9 trials; $q = 1.25\times10^{-8}$) and from 53.48% (no trial average; $q = 0.067$) to 60.72% (average of 9 trials; $q = 4.99\times10^{-8}$), respectively. These results suggest that the EEG DNN succeeded to decode sub-categorization of visual objects while the fMRI DNN did not.

Performance distribution and the number of significant models were further analyzed to confirm the robustness of the results. First, performance distribution of nine models for categorization statistically reached above chance when trials were not averaged (57.07 ± 0.15%; t-test against chance level 50%, $t = 47.33$, $P = 4.39\times10^{-11}$,

FDR corrected $q < 0.05$) (Figure 4D). The performance improved as a function of the number of averaged trials (linear regression, $R^2 = 0.87$, slope coefficient, $t = 16.90$, $P = 1.34\times10^{-20}$), resulting in 73.41 ± 0.74% with an average of nine trials ($t = 31.73$, $P = 1.06\times10^{-9}$). Decoding performance for both face and object sub-categorization also improved as the number of averaged trials increased (face sub-categorization: $R^2 = 0.44$, slope coefficient, $t = 5.78$, $P = 7.51\times10^{-7}$; object sub-categorization: $R^2 = 0.59$, slope coefficient, $t = 8.00$, $P = 4.78\times10^{-10}$). As for male face/female face classification, the mean accuracy in the condition where trials were not averaged was significantly above chance (51.90 ± 0.60%; $t = 3.19$, $P = 0.0129$, FDR corrected). When nine trials were averaged, the accuracy reached 57.51 ± 0.64% ($t = 11.76$, $P = 2.49\times10^{-6}$). Notably, the highest accuracy was 59.46% in the best performing model. As for natural object/artificial object classification, the accuracy improved from 52.91 ± 0.60% (no trial average; $t = 4.86$, $P = 0.0013$, FDR corrected) to 58.71 ± 0.62% (average of 9 trials, $t = 14.08$, $P = 6.30\times10^{-7}$). The highest accuracy was 61.56% in the best performing model.

Second, while all nine models were significant in categorical classification across all averaged trial conditions, the number of models with significant sub-categorical classification increased as a function of the number of averaged trials (binomial test, threshold at $P = 0.05$, FDR corrected). In face sub-categorization, the number of significant models increased from one to nine, whereas in object sub-categorization, the number increased from two to nine.

Visualization by guided Grad-CAM revealed that our DNN model valued differential periods between categorization and sub-categorization (Figure 4E). The model started to value the timepoint at 120 ms and 336 ms in the face and object trials, respectively ("emergence latency"; see Materials and Methods for details). Regarding sub-categorization, the four emergence latencies were 788 ms for male face trials, 216 ms for female face trials, 288 ms for natural object trials, and 328 ms for artificial object trials, all of which were slower than that for face trials. Meanwhile, emergence latencies for female face and natural object trials were longer than those for object trials. This "categorization-then-subcategorization" tendency in the latency of first valued timepoints was also observed in the latency with the maximum guided Grad-CAM value ("maximum latency") (Figure 4F). The maximum latency varied among trials, and their ranges differed between categorization and sub-categorization: the maximum latency in categorization varied between 104 ms and 912 ms (median = 498 ms), while that in face sub-categorization and object sub-categorization varied between 236 ms and 960 ms (median = 556 ms) and between 228 ms and 932 ms (median = 541 ms),

respectively. The maximum latency in categorization was significantly shorter than that in the face and object sub-categorization (Kruskal–Wallis test, $\chi^2 = 9.70$, $P = 0.0078$; post-hoc Tukey–Kramer test, $P < 0.05$). Notably, the maximum guided Grad-CAM values in each trial type were significantly larger than those in non-learned models (Figure 4G; Wilcoxon signed rank test against guided Grad-CAM values in non-learned models, FDR corrected $q < 0.05$). When grouping guided Grad-CAM values into six brain regions by locations of electrode channels, all regions showed significant guided Grad-CAM values (Figure 4 – figure supplement 2). Of note, the highest value of guided Grad-CAM in categorization was in the occipital cortex, while the frontal and parietal cortices were more valued in sub-categorization (Friedman's test with post-hoc Tukey-Kramer method, $P < 0.05$).

Collectively, these results indicate that our EEG DNN model successfully captured not only the categorical representation, but also the sub-categorical representation of visual objects in the temporal domain of neural responses. Our results experimentally support that the visual object categorization is processed in a coarse-to-fine manner in the temporal axis.

Concurrent neural decoding by combined DNN model

Simultaneous acquisition of fMRI and EEG data enables concurrent neural decoding using an integrated DNN. Thus, we performed neural decoding for concurrent fMRI and EEG data with the combined fMRI–EEG DNN (Figure 5A). In the DNN, feature extraction layers, 3D convolution layers for fMRI data and temporal blocks for EEG data, were finally concatenated and fully-connected.

When the fMRI and EEG activities of nine trials were averaged, ensemble performance of the fMRI–EEG DNN was 90.05%, 60.70%, and 63.64% for categorization, face sub-categorization, and object sub-categorization, respectively, all of which were significantly above the chance level (Figure 5B; binomial test, $P = 3.03\times10^{-240}$, $P = 2.51\times10^{-9}$, and $P = 2.44\times10^{-13}$). Distributions of performance in the nine models were also significant against chance level (categorical, 87.49 ± 1.10%, $t = 78.97$, $P = 7.37\times10^{-13}$; face sub-categorical, 58.40 ± 1.43%, $t = 40.84$, $P = 1.42\times10^{-10}$; object sub-categorical, 59.97 ± 2.36%, $t = 25.44$, $P = 6.11\times10^{-9}$; all $q < 0.05$ in FDR correction). The best performing models showed performances of 90.46%, 63.81%, and 70.04%. A binomial test revealed that nine, seven, and eight models had significant decoding performance for categorization, face sub-categorization, and object sub-categorization, respectively (threshold at $P = 0.05$, FDR corrected).

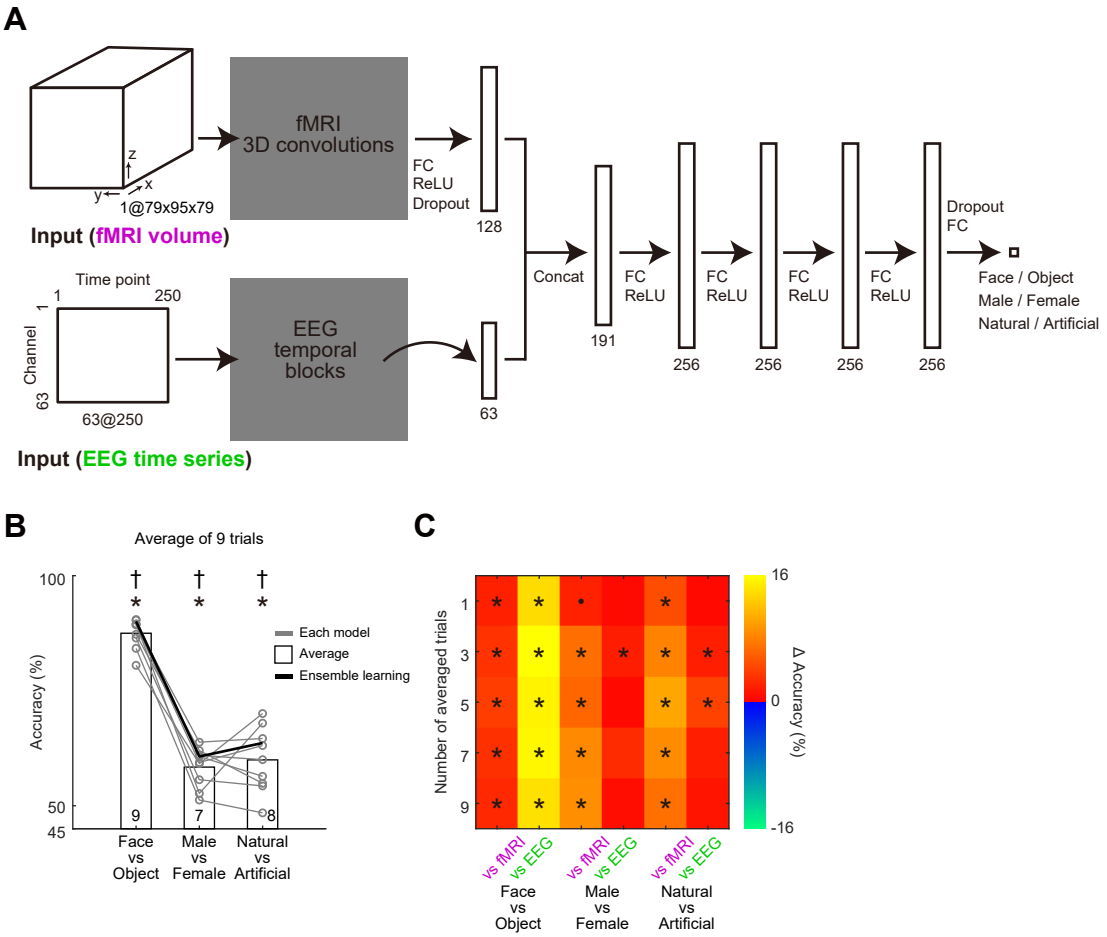
A
z
x
y
1@79x95x79
Input (fMRI volume)
fMRI
3D convolutions
FC
ReLU
Dropout
128
Time point
1
250
Channel
1
63
63@250
Input (EEG time series)
EEG
temporal
blocks
63
Concat
191
FC
ReLU
256
256
256
256
Dropout
FC
Face / Object
Male / Female
Natural / Artificial
B
Average of 9 trials
100
50
45
Accuracy (%)
Each model
Average
Ensemble learning
9
7
8
Face
vs
Object
Male
vs
Female
Natural
vs
Artificial
C
Number of averaged trials
1
3
5
7
9
vs fMRI
vs EEG
Face
vs
Object
Male
vs
Female
Natural
vs
Artificial
16
0
-16
Δ Accuracy (%)

Figure 5

Figure 5. Combined fMRI–EEG DNN model. (A) Architecture of the model. First, fMRI volume data and EEG time series data are separately convolved by conv3Ds and temporal blocks shown in Figure 2. Next, resultant outputs are concatenated, which are then processed using four fully connected layers. **(B)** Generalization decoding performance for the test dataset (trial average of nine). Bar denotes mean accuracy across models. Circles, accuracy in each model. *, binomial test for ensemble accuracy against chance level (50%), FDR corrected $q < 0.05$. †, t-test for accuracy distribution against chance level, FDR corrected $q < 0.05$. The number of significant models (binomial test, $q < 0.05$) is shown above the abscissa. **(C)** Improvement in decoding accuracy of fMRI–EEG DNN against accuracy in the fMRI DNN model and the EEG DNN model. Differences in accuracy are color-coded. Note that none of the fMRI DNN or the EEG DNN shows higher accuracy than the fMRI–EEG DNN. *, t-test, FDR corrected $q < 0.05$. •, $P < 0.05$, uncorrected.

In the comparison of the classification accuracy of the combined fMRI–EEG DNN with accuracy for fMRI data alone (Figure 3) or EEG data alone (Figure 4), accuracies in the fMRI–EEG DNN were higher than those in the fMRI DNN and the EEG DNN in all averaged trial conditions (Figure 5C). Categorical classification accuracy of the fMRI–EEG DNN was significantly increased in all averaged trial conditions (t-test, FDR corrected $q < 0.05$). Regarding sub-categorical classification, the fMRI–EEG DNN showed significantly higher performance than did the fMRI DNN in all averaged trial conditions ($q < 0.05$). However, the difference in performance between the fMRI–EEG DNN and the EEG DNN did not reach statistical significance except for when three trials were averaged. These results suggest that concurrent usage of spatial and temporal neural data contributes to improve neural decoding of visual objects.

Discussion

This study found that the categorical representation of visual objects is supported by spatially-organized brain-wide activations in a coarse-to-fine manner. To our knowledge, the present study is the first to examine neural representation of visual objects using multimodal DNNs. This approach enabled highly resolved brain-wide spatiotemporal neural decoding. The present findings would contribute to construct a novel paradigm for trial-based neural decoding of internal processes in patients with impaired consciousness in clinical practice.

DNNs, which achieved state-of-the-art results for the large-scale image recognition (Simonyan and Zisserman, 2015), is spreading to apply to neural data, known as neural decoding (Hong et al., 2016; Horikawa and Kamitani, 2017; Yamins and DiCarlo, 2016). In fMRI DNN, we managed to use 3D voxel data as an input, while some previous studies used 2D slice data (Sarraf et al., 2017) or 2D flattened cortical data as an input (Tsumura et al., 2021b) (see also other studies where 3D voxel data were used as input data (Frey et al., 2021; Vu et al., 2020; Wang et al., 2020; Zhao et al., 2018)). Such image-like input could be learned by DNNs that have already been trained with a large-scale image set (e.g., ImageNet), such as VGG. The re-training approaches such as fine-tuning or transfer learning help improve decoding performance. Given that trained models for spatio-temporal data, which are suitable for applying to concurrent fMRI and EEG data, are currently unavailable, we could not employ such a strategy in this study. However, there are several clear advantages of our approach, such as (1) inclusion of neural activities in the subcortical structures, which are ignored in the 2D flattened map, and (2) coverage of whole brain structures, which are impossible in 2D slice data. Indeed, we could identify the involvement of the amygdala and thalamus in categorical classification of visual objects. Accumulating open 3D-voxel data available for deep learning is promising for further improving in 3D fMRI decoding using DNNs.

In EEG DNN, we employed a TCN (Bai *et al.*, 2018) in which time-series data are convoluted in a dilated manner. The TCN has excellent performance for non-neural data such as character-level and word-level language modeling. Successful application of the TCN to EEG data in the present study would expand the possibility for the application of the TCN to other biological data such as electromyography.

We obtained high decoding performance for both categorical and sub-categorical classification (*c.f.,* in the fMRI–EEG DNN, ensemble performance reached 90.05%, 60.70%, and 63.64% for categorical and two sub-categorical classifications, and the best performing model showed 90.46%, 63.81%, and 70.04%). It is difficult to

directly compare the performance in this study with that in other studies because experimental paradigms (e.g., decoding technique, stimulus, and sample size) are quite different. Nevertheless, the present finding that the fMRI–EEG DNN outperformed the fMRI DNN and the EEG DNN indicates that a larger number of modal data would improve the precise identification of the neural representation of visual objects.

In machine learning including deep learning, learning performance is usually evaluated once for the test dataset (Raschka, 2020). This study employed the ensemble learning approach (Rokach, 2009) in which obtained classification labels from nine models were voted and the majority label was treated as the predicted label for the trial. We confirmed the robustness of this ensemble approach using two distinct indices, performance distribution and the number of significant models. Considering the inherent properties of neural data as noisy and variable between trials, this combinational approach is suitable for the application of deep learning to neural data. For example, when comparing an fMRI volume or an EEG time series with pictures in ImageNet, neural data have some background noise, c.f., we can hardly detect the internal process of the subject by glimpsing an fMRI volume or an EEG time series. In addition, it is difficult to collect adequate amount of neural data that are comparable to the number of pictures in ImageNet (>10,000,000). The noisiness and smaller sample size would result in higher variance in the decoding performance. The ensemble learning approach would reduce the variance of classification errors, and in addition to a point estimate of the classification performance, our approach can also evaluate the variance of performance across models.

Brain regions where the fMRI DNN valued for categorization closely matched with regions identified in the fMRI univariate analysis, including the ventral temporal cortex (FFA for face stimuli and LOC for object stimuli), consistent with previous findings (Haxby *et al.*, 2001; Kanwisher *et al.*, 1997; Kourtzi and Kanwisher, 2000; Malach *et al.*, 1995). Notably, we revealed the involvement of the prefrontal cortex and amygdala, which have been also demonstrated in monkey (Freedman et al., 2001) and human (Kreiman et al., 2000) electrophysiology. Considering the wider distribution of significant voxels in the guided Grad-CAM than that in the univariate analysis, deep learning techniques would have greater sensitivity to detect responsible brain regions for visual object categorization.

Fine and coarse information processing in visual object categorization is still controversial (Carlson *et al.*, 2013; Cichy *et al.*, 2014; Dobs *et al.*, 2019; Liu *et al.*, 2002; Mack and Palmeri, 2011). Our deep learning-based approach provided robust evidence to this issue: temporal trajectory of neural decoding by the EEG DNN

demonstrated that the neural activities first conveyed categorical information, followed by sub-categorical information. These results are in line with the theoretical model of "coarse-to-fine" categorization (Bar, 2004; Rousselet et al., 2002; Serre et al., 2007). One notable observation in this study is that peak time in the guided Grad-CAM value (maximum latency) varied among trials, which could be possibly caused by temporal variability of behavioral states (e.g., attention) among trials. EEG DNN could detect such variability at any time point owing to shift invariance in the temporal domain by dilated convolutions. Trial-dependent latencies of category/sub-category information in the neural responses would be related to the persistent time windows with significant object decoding reported in other studies (Cichy *et al.*, 2014; Dobs *et al.*, 2019).

Compared with the situation in which fMRI and EEG recordings are separately conducted, the concurrent fMRI/EEG recording has several advantages. First, the variance of behavioral outcomes such as the correct performance and reaction time could be lower than those obtained in separate sessions or days. Second, concurrent usage of fMRI and EEG data can result in better decoding performance than the usage of fMRI alone or EEG alone, as shown in this study. It is promising to add other bio-signals for detecting other aspects of information of visual objects such as valence or arousal.

An advanced M/EEG–fMRI fusion approach has recently attracted attention as it links multivariate neural response patterns based on representational similarity (Cichy and Oliva, 2020; Cichy *et al.*, 2014) while M/EEG and fMRI data are acquired on separate sessions. The combination of concurrent multimodal recording and representational similarity-based fusion approach would be a useful method to more accurately elucidate neural representation in the sensory system.

Materials and Methods

Participants

A total of 53 healthy right-handed subjects (34 males and 19 females; age, 18–25 years) with normal or corrected-to-normal vision participated in the study. Of them, data from 3 participants were excluded due to low behavioral performance: 2 participants had low correct performance (< 80% at least for one sub-category) and 1 participant had slow responses (mean response latency > 1.5 s). Participants received 1,000 yen per hour for each session. The target number of participants was determined based on the number of trials required in previous deep learning research in neuroscience (e.g., Supratak et al (Supratak et al., 2017)) before starting the collection of neuroimaging data. This study was approved by the institutional review board of Kochi University of Technology, Kami, Japan and complied with the Declaration of Helsinki. Written informed consent was obtained from all participants.

Experimental design

Stimuli. Stimuli consisted of color images (size, 7.4° × 5.8°) and a white fixation cross (size, 0.3° × 0.3°). Forty color images from face and inanimate object categories were used as experimental stimuli. The 20 face and 20 object images were selected from ATR Facial Expression Database DB99 (http://www.atr-p.com/products/face-db.html) and ALOI (Amsterdam Library of Object Images; https://aloi.science.uva.nl/) (Geusebroek et al., 2005), respectively. Facial images were obtained from 4 identities: two male individuals (male face sub-category) and two female individuals (female face sub-category), with neutral facial expressions and 5 different angles. The object images were obtained from 4 identities: two natural objects (paprika and broccoli; natural object sub-category) and two artificial objects (hat and whistle; artificial object sub-category), with 5 different angles.

All images were contrast- and luminance-matched using in-house MATLAB (MathWorks, Natick, MA) code (Keerativittayayut et al., 2018) adapted from SHINE toolbox (*lumMatch* function) (Willenbockel et al., 2010) for each RGB channel, after figure-ground separation. Based on a prior study (Stigliani *et al.*, 2015), we evaluated (1) contrast, (2) luminance, (3) similarity to other stimuli of the same category, (4) visual field coverage, and (5) spatial frequency power distributions, after converting RGB color images to grayscale images by *rgb2gray* function in MATLAB (Figure 1 – figure supplement 1). In addition, hue histograms of color images were also evaluated for each sub-category.

Task procedure. During the recording of brain activities by concurrent fMRI and EEG, participants performed a visual object classification task, in which sequences of experimental stimuli were presented. Participants judged the corresponding visual sub-categories (male face, female face, natural object, and artificial object) and pressed a corresponding button (Figure 1B–C). At the beginning of each scanning run, a task instruction was presented that indicated the correspondence between 4 positions of buttons and 4 sub-categories. To exclude the possibility that the learning of stimulus-response mapping would affect neural activities, the correspondence was pseudorandomized across runs. In the scanning run, a fixation cross was first presented for 30 s, followed by a sequence of 50 trials. In 40 trials, face or object images were presented, and in 10 trials, blank images were presented to improve dissociation of hemodynamic responses between trials. In each trial, a stimulus was presented for 500 ms, followed by a fixation cross for 9,500 ms. Thus, one scanning run took a total of 530 s. All stimulus images were presented once in each scanning run, and the order of stimulus presentation was pseudorandomized. No visual feedback was presented. Participants practiced the task before starting concurrent fMRI and EEG.

Participants conducted 9.96 ± 2.88 runs (mean ± SD; total 496 runs) while simultaneous fMRI and EEG recordings were performed. Neural data were accordingly obtained from a total of 24,800 trials (9,920 face trials, 9,920 object trials, and 4,960 blank trials). Trials wherein participants did not respond within 2 s (1,005 trials, 5% of total face/object trials) were excluded. Correct performance and reaction time were calculated for each category and sub-category (Figure 1D–E). Only data from correct trials were used for the succeeding analyses (18,007 face and object trials). The task was programmed and administered using Psychtoolbox version 3 (http://psychtoolbox.org/). The visual stimuli were projected on a screen located behind the scanner using PROPixx DLP LED projector (VPixx Technologies, Canada). Participants viewed the projected visual stimuli through a mirror attached to a head coil.

fMRI procedure and preprocessing

MRI scanning was performed using a 3T MRI scanner (Siemens Prisma, Germany) with a 64-channel head coil. Functional images were acquired using an echo-planar imaging sequence [repetition time (TR): 2.0 s; echo time (TE): 27 ms; flip angle (FA): 90 deg; 36 slices; slice thickness: 3 mm; in-plane resolution: 2 × 2 mm]. Multiband accelerated imaging was unavailable because the guideline for the simultaneous fMRI–EEG recoding by the developer of the MRI-compatible EEG system did not permit it (Brain

Products, Germany). Each functional run involved 265 volume acquisitions. The first 5 volumes were discarded for analyses to take into account the equilibrium of longitudinal magnetization. High-resolution anatomical images were acquired using an MP-RAGE T1-weighted sequence (TR: 2,400 ms; TE = 2.32 ms; FA: 8 deg; 192 slices; slice thickness: 0.9 mm; in-plane resolution: 0.93 × 0.93 mm).

All preprocessing was performed using SPM12 software (http://fil.ion.ac.uk/spm/). Functional images were first temporally realigned across volumes and runs, and the anatomical image was co-registered to a mean image of the functional images. The functional images were then spatially normalized to a standard Montreal Neurological Institute (MNI) template, with normalization parameters estimated for the anatomical scans. The images were resampled into 2-mm isotropic voxels. For the univariate fMRI analysis, functional volumes were spatially smoothed with a 6-mm full-width at half-maximum Gaussian kernel. For the deep learning analysis, we did not apply spatial smoothing to fMRI data.

EEG procedure and preprocessing

High-density EEG data were acquired from MRI-compatible 64 sintered Ag/AgCl electrodes with extended scalp coverage, including 63 scalp recording electrodes and an electrode on the upper back for electrocardiogram (ECG) (BrainCap-MR, Brain Products, Germany). The electrodes AFz and FCz served as ground and online reference, respectively. All signals were recorded with a bandpass of 0.016–250 Hz and digitized at 5,000 Hz (BrainAmp MR Plus, Brain Products). Electrode impedances were lowered to <10 kΩ before recording and monitored throughout the experiment.

Offline preprocessing was firstly performed in Brain Vision Analyzer 2 (Brain Products). First, we removed the MR gradient artifact using an average template subtraction method (Allen et al., 2000) and down-sampled the resultant EEG data to 250 Hz. Second, the EEG data were bandpass filtered between 1 Hz and 100 Hz. Third, the cardiovascular artifact was removed from the EEG data. Heartbeat detection using the ECG channel was then performed using a semiautomatic template-matching procedure. Fourth, eye blink correction was performed using independent component analysis (ICA). Since specific sensors for vertical/horizontal electrooculography (EOG) were not used, we used the EEG data obtained from electrode FP1 as a substitute for the vertical activity of the EOG channel. Finally, we applied ICA-based denoising techniques implemented in EEGLAB (*runica* function). Then, each independent component was labeled using ICLabel plugin (https://github.com/sccn/ICLabel). Components with <75% probability of being in the brain signal were flagged as artifacts and rejected.

The resultant artifact-removed EEG data were further processed in EEGLAB as follows. If the signal was continuously noisy at a certain channel in a scanning run, the EEG data were interpolated at the corresponding channel by replacing it with averaged EEG data from the neighboring channels. Accordingly, 16 out of 3,150 channels (50 participants × 63 scalp channels) were interpolated. Thereafter, the EEG data were segmented into 1.5-s epochs (0.5 s pre-stimulus to 1.0 s post-stimulus) time locked to the stimulus onset and were baseline-normalized. The data were then inspected manually to detect any trials containing any excessive noises. After excluding 742 trials, 17,265 trials were included in the analyses. Data were then digitally re-referenced to the average of all scalp channels.

For the ERPs, we first trial-averaged EEG time series for each EEG channel separately. Then, the resultant ERPs were grand-averaged across participants to obtain population ERPs for each EEG channel. Spectral power was also calculated using *mtspecgramc* function in Chronux toolbox for MATLAB (http://chronux.org/). The parameters for multi-taper time-frequency spectrum were as follows: window = 200 ms, step size = 80 ms, time-bandwidth product = 3, and number of tapers = 5.

Respiration and eye-tracking recording

Respiratory data were recorded by using the BIOPAC MP160 system with AcqKnowledge software (BIOPAC Systems, Inc., Goleta, CA). Raw physiological time series of respiration were sampled at 2,000 Hz and then were preprocessed in PhysIO toolbox (Kasper et al., 2017) for MATLAB as follows. First, the respiration data were down-sampled to 250 Hz and bandpass-filtered between 0.1 Hz and 5.0 Hz. Then, both the respiratory phase and respiratory volume per time (RVT) for the respiratory data were obtained. From these physiological determinants with noise modeling, (1) 8 RETROICOR (RETROspective Image CORrection) regressors were computed via a 4th order Fourier expansion of respiratory phase (Harvey et al., 2008), and (2) respiratory response was calculated by convolution of RVT with the respiration response function (Birn et al., 2006). Accordingly, we obtained a total of 9 regressors and treated them as nuisance regressors in the fMRI univariate analysis (see "*fMRI univariate analysis*" section below).

To check whether participants fixated the eye position, the eye position during task runs was monitored online at a sampling rate of 1,000 Hz using an MRI-compatible infrared camera-based eye tracker (EyeLink 1000, SR Research, Ontario, Canada).

Deep neural network classifier

To explore spatiotemporal neural activity representing visual object category and sub-category, in-house deep neural network (DNN) classifiers (Simonyan and Zisserman, 2015) were constructed (Figure 2). The architecture of the models is described below. The model training, validation, and testing framework for binary classification (face/object, male face/female face, or natural object/artificial object images) were implemented using PyTorch (https://pytorch.org/). The scripts used for network modeling and analyses are available in GitHub (https://github.com/masaki-takeda/dld).

The whole dataset included 17,265 trial-based concurrent fMRI–EEG data from 50 participants. Data from 45 participants were used as a training/validation dataset for optimizing hyperparameters, and independent data from 5 participants were used as a hold-out test dataset. In the training/validation dataset, data from 5 participants were treated as a unit of the fold, resulting in 9-fold datasets (Figure 2B). After preliminarily determining network architectures, we conducted grid searches for hyperparameter tuning by using the train/validation dataset (see "*Grid search*" section). After completing hyperparameter tuning, we next evaluated the generalization of neural decoding performance for the independent test dataset: 9-fold cross-validation was performed for the training/validation dataset, and then the resultant 9 learned models were applied to the hold-out test dataset. For each binary classification for the test dataset, we equalized the number of trials for face/object, male face/female face, and natural object/artificial object by excluding a surplus of trials in the test dataset. This resulted in the chance level of 50% (categorization: 734 face trials vs 734 object trials; face sub-categorization: 386 male face trials vs 386 female face trials; object sub-categorization: 359 natural object trials vs 359 artificial object trials). The decoding performance for visual object categorization and sub-categorization was then investigated using the following network models based on fMRI data alone (Figure 2C), EEG data alone (Figure 2D–E), and combined fMRI–EEG data (Figure 5A).

fMRI DNN. The current DNN model for fMRI data (fMRI DNN) consisted of a three-dimensional (3D) convolutional neural network (Frey *et al.*, 2021) (Figure 2C). The 3D CNN included three convolution layers (*conv3d* function) for extracting spatial features and two fully connected layers. We also examined other networks with a greater number of convolution and fully connected layers, but decoding performance did not change drastically. Based on the preliminary test for the training/validation dataset, the following hyperparameters and settings were selected: batch size of 100, early stopping,

and Adam optimizer (Kingma and Ba, 2015). We then grid-searched for the number of averaged trials, learning rate, and weight decay.

The input datasets were trial-based non-smoothed fMRI volumes and were prepared as follows. To consider the delayed hemodynamic response of the BOLD signal, an fMRI volume that was acquired at 4 s after the onset of the visual stimulus presentation was collected (Figure 2A). The input fMRI volume was spatially normalized to standard MNI space, and signal intensities outside of the brain were masked to zero. Then, the signal intensity in each voxel was z-scored by the mean and SD values of the signal intensities in the volume.

EEG DNN. Seven DNN models for EEG data (EEG DNN) were tested in the preliminary inquiry for training/validation dataset: (1) a custom DNN model including four *conv1d* layers, (2) a custom DNN model including four *conv2d* layers for band-passed EEG time series (delta, theta, alpha, beta, and gamma frequency band × time), (3) a recurrent neural network (RNN) model using long short term memory (LSTM), (4) a convolutional RNN (ConvRNN) model using LSTM after two *conv1d* layers, (5) a spatial-temporal neural network originally designed for P300 detection (Zhang et al., 2021), (6) a generic temporal convolutional network (TCN) (Bai *et al.*, 2018), and (7) a modified version of the TCN. Among them, we adopted the TCN model (6) because of its superior decoding performance and its simplicity (Figure 2D–E). The TCN model has several distinguishing characteristics such as the dilated convolutions (dilated *conv1d*) and the causal architecture (no information leakage from future to past). Based on the preliminary test, a batch size of 100 and early stopping were selected. We then grid-searched for the number of averaged trials, learning rate, and kernel size. The dilation factors for each kernel size were determined to cover all the input time series.

The input datasets were trial-based EEG time series data and were prepared as follows. Preprocessed EEG time series data were collected from 63 scalp channels. The segmented data obtained at 1 s from the stimulus onset (250 samples at a sampling rate of 250 Hz) was used (Figure 2A). The resultant segmented data were further z-scored by the mean and SD values of the signal amplitude for each channel.

DNN for combined data. We also tested whether concurrent usage of both fMRI and EEG data improves the decoding performance relative to the usage of fMRI data alone or EEG data alone. The DNN architecture for combined data (fMRI–EEG DNN) is shown in Figure 5A. In the model, the feature extraction layers of the two networks for fMRI and EEG data were concatenated and then fully connected. Both fMRI and EEG

networks were pre-trained separately; thereafter the initial parameters of convolution layers in the fMRI–EEG DNN were set to the pre-trained parameters.

Grid search. Before conducting the grid search, we preliminarily searched several hyperparameters such as the number of layers and kernel sizes, and found that several hyperparameters affected the decoding performance: the number of averaged trials, weight decay, and learning rate for fMRI data, and the number of averaged trials, kernel size for delated convolutional layers, and learning rate for EEG data. Thus, the grid searches were conducted to obtain the optimal combination of these hyperparameters. For the fMRI DNN, the following hyperparameters were tuned: the number of averaged trials (1, 3, 5, 7, 9), learning rate (0.0001–0.01 in 5 divisions), and weight decay (0–0.01 in 5 divisions). For the EEG DNN, the following hyperparameters were tuned: the number of averaged trials (1, 3, 5, 7, 9), learning rate (0.0001–0.01 in 5 divisions), and kernel size (2–9 in 5 divisions). For the combined fMRI–EEG DNN, the following hyperparameters were tuned: the number of averaged trials (1, 3, 5, 7, 9), learning rate of feature extraction layers for fMRI data (0.001 or 0.0001), weight decay for fMRI data (0 or 0.001), learning rate of feature extraction layers for EEG data (0.0001 or 0.0005), and kernel size for EEG data (2, 5, 7, 9). For the combined fMRI–EEG DNN, we fixed the learning rate and weight decay for fully connected layers as 0.001 and 0, respectively. All grid search analyses were conducted using the training/validation dataset, and the hold-out test dataset was completely free from the hyperparameter settings.

Data averaging. A single-trial neural signal contains noises derived from the brain and measurement hardware, and one of the solutions to increase the signal to noise (S/N) ratio is to average the data. To estimate how the S/N ratio of brain signals affects the decoding performance, we tested the decoding using both data from a single trial data and averaged data (number of averaged trials ≤ 9). We averaged fMRI or EEG data within each subject and then augmented the data to match the total number of averaged data with the number of all trials (e.g., averaged data from 9 trials were augmented by a factor of 9).

Visualization of relevant neural activity in DNNs. To highlight important spatiotemporal neural activities for visual object classification, visualization approaches in deep learning were applied. For both the fMRI DNN and EEG DNN, guided Gradient-weighted Class Activation Mapping (guided Grad-CAM) (Selvaraju *et al.*, 2019) was

used to create a high-resolution class-discriminative visualization. Guided Grad-CAM was created by multiplying guided backpropagation at the input of the first convolution layer with Grad-CAM at the last convolution layer. The original Grad-CAM uses global average pooling of the gradients (Selvaraju *et al.*, 2019), but we did not use it like Layer-CAM (Jiang et al., 2021), because the smoothing effect of pooling was extremely strong, especially for the 3D convolution. In the end, we obtained multiple guided Grad-CAM maps, the number of which corresponded to the number of trials. The obtained guided Grad-CAM values were not normalized. As for EEG guided Grad-CAM, the model without residual connections was used because those connections can cause gridding artifacts with the dilated convolution layers and obfuscate the visualization results (Yu et al., 2017). The resultant guided Grad-CAM values in the EEG DNN were then smoothed using the Gaussian-weighted moving average with a 40-ms window.

Statistics

General. All statistical analyses were performed in MATLAB and Python. All data are shown as mean ± standard error of the mean unless otherwise stated. Sample sizes were not based on a priori power calculations but are comparable to other studies in the field using similar techniques (Supratak *et al.*, 2017).

fMRI univariate analysis. The fMRI univariate analysis was conducted as in previous studies (Matsui et al., 2022; Tanaka et al., 2020; Tsumura et al., 2021a; Tsumura *et al.*, 2021b) (Figure 3A–C). In the first-level analysis, a GLM (Worsley and Friston, 1995) approach was used to estimate parameter values for task events. The events of interest were the presentation of male face, female face, natural object, artificial object stimuli (4 regressors). Only correct trials were coded as regressors. Those task events were time locked to the onset of stimulus presentation and then convolved with a canonical hemodynamic response function implemented in SPM. Additionally, the onset of blank stimuli, six-axis head movement parameters, white matter signal, CSF signal, and effect of respiration (9 regressors) were included in GLM as nuisance effects, totaling to 18 nuisance regressors. Then, parameters were estimated for each voxel across the whole brain.

In the group level analysis, maps of parameter estimates were first contrasted within individual participants. The contrasted maps were then collected from participants and subjected to group-mean tests. Voxel clusters were first identified using a voxel-wise uncorrected threshold of $P < 0.001$. The voxel clusters were then tested for a significance across the whole brain with a threshold of $P < 0.05$ corrected by family-

wise error (FWE) rate based on permutation methods (5,000 permutations) implemented in the *randomise* function in FSL suite (http://fmrib.ox.ac.uk/fsl/), which was empirically validated to appropriately control the false positive rate in a previous study (Eklund et al., 2016). Peaks of significant clusters were then identified and listed in tables. If multiple peaks were identified within 12 mm, the most significant peak was retained.

EEG univariate analysis. For each participant, a single ERP and power spectrogram were constructed by averaging data across all trials separately for each visual category or sub-category in each channel. We then averaged ERPs/spectrograms from all participants, which resulted in a grand-averaged of 63-ch ERPs/spectrograms for face, object, male face, female face, natural object, and artificial object trials (Figure 4A–C; Figure 4 – figure supplement 1A). The subsequently obtained spectrograms were further bandpass-filtered to divide them into alpha (8–15 Hz), beta (15–30 Hz), and gamma (30–70 Hz) frequency bands.

For statistical inference on differential EEG time series of ERP and power spectrogram (e.g., face minus object trials) against zero, permutation-based cluster-size inference (Dobs *et al.*, 2019; Maris and Oostenveld, 2007) (i.e., a cluster refers to a set of contiguous time points) was performed separately for each channel. Time points that exceeded the 97.5th percentile of the permutation distribution (N = 5,000) served as cluster-inducing time points (i.e., equivalent to $P < 0.05$, two-sided). The significance of clusters was set at the threshold of $P < 0.05$ (Figure 4 – figure supplement 1B–D).

Neural decoding performance. We trained our DNN models using 9-fold cross-validation with the training/validation dataset and then evaluated their generalization performance with the test dataset. Thus, a total of 9 generalization performances were obtained for the fMRI DNN model, EEG DNN model, and combined fMRI–EEG DNN model. Three different statistical tests were performed to examine the significance of decoding accuracies and their robustness (Figure 3D, 4D, and 5B). First, an ensemble learning method with majority voting (Rokach, 2009) was conducted. In this scheme, each base model (N = 9) made a prediction and voted for each sample (a stimulus in a trial). Only the sample class with the highest votes was included in the final predictive class. Statistical significance was tested using the binomial test against chance level (50%). Second, the t-test was used for distribution of accuracies obtained by the nine models against chance level, with FDR corrections ($q < 0.05$). Third, the binomial test

was used for each model against chance level, with FDR corrections ($q < 0.05$), and the number of significant models was counted.

Visualization of neural decoding. The model that showed maximum accuracy in the average of nine trials was selected for the succeeding visualization analysis.

Statistical significance of guided Grad-CAM for fMRI data was evaluated by permutation-based cluster-size inference (Dobs *et al.*, 2019; Maris and Oostenveld, 2007) used in the present univariate analyses (Figure 3E; see "*fMRI univariate analysis*" section). The spatial distribution of the significant guided Grad-CAM was also evaluated based on the Harvard-Oxford cortical and subcortical structural atlases, which covered 48 cortical and 21 subcortical structural areas, implemented in FSL (Figure 3 – figure supplement 4; *HarvardOxford-cort-maxprob-thr50-2mm* and *HarvardOxford-sub-maxprob-thr50-2mm*). Regarding cortical areas, we divided them into areas in the right and left hemispheres, and obtained 95 cortical regions (note no voxels for left supracalcarine cortex). Regarding subcortical areas, we bilaterally removed areas labeled as the cerebral white matter, lateral ventricle and cerebral cortex, and obtained 15 subcortical regions. We examined whether guided Grad-CAM values in each brain regions ($N = 110$) were significantly higher than values obtained when DNNs were not learned. In the non-learned model, weight values were initialized and fixed as random values. Statistical significance was evaluated by Wilcoxon signed rank test with FDR corrections ($q < 0.05$).

We also conducted two types of correlation analysis (Spearman's rank correlation; Figure 3 – figure supplement 5) and three types of ROI analysis to examine the spatial overlap between activations in the univariate analysis and guided Grad-CAM (Figure 3F, Figure 3 – figure supplement 6). The correlation was calculated either for the voxels masked by FSL's gray matter prior volume (*avg152T1_gray* template, threshold = 0.5) or for the voxels masked by the maps of meta-analysis obtained from Neurosynth (Yarkoni *et al.*, 2011) (https://neurosynth.org) using the terms "face" and "object" (uniformity test). In the ROI analysis, we examined whether guided Grad-CAM values in each ROI were significantly higher than the values obtained using the non-learned DNNs. ROIs were selected based on (1) the univariate analysis (Figure 3 – figure supplement 1; Figure 3F), on (2) the anatomical structures from the Harvard–Oxford atlas (Figure 3 – figure supplement 6A), or on (3) the maps of meta-analysis obtained from Neurosynth (Figure 3 – figure supplement 6B). In the Harvard-Oxford atlas (2), we created ROIs, in which the peak voxel of clusters identified by the univariate analysis was included, by thresholding the probabilistic map of the region by

50%. Note that we combined MFGi and MFGs for object trials identified in the univariate analysis into one ROI as MFG. In the Neurosynth (3), ROIs were selected with the uniformity test using the terms “face” and “object.” ROIs with more than 50 voxels were used in the analysis. In all ROI analyses, statistical significance was evaluated using Wilcoxon signed rank test with FDR corrections ($q < 0.05$).

Guided Grad-CAM for EEG data were quantified according to two indices, emergence latency and maximum latency. First, guided Grad-CAM values for each trial were averaged across 63 channels. Then, for each trial type, we defined the emergence latency as the time point where the value first exceeded the level at 20% of the maximum of channel-averaged guided Grad-CAM values (Figure 4E). We obtained similar results when the threshold was changed to 10%, 15%, 25%, and 30%. The maximum latency for each trial was also defined as the time point where the guided Grad-CAM showed maximum value in each trial (Figure 4F), and statistical significance in the maximum latency across categorization/sub-categorization was tested using Kruskal–Wallis test with post-hoc Tukey-Kramer test ($P < 0.05$). We also calculated the maximal latencies and corresponding guided Grad-CAM values when the DNN was not learned. Then, we examined whether guided Grad-CAM values at the maximum latency were significantly higher in the learned model than those in the non-learned model using Wilcoxon signed rank test with FDR corrections ($q < 0.05$; Figure 4G). Finally, we tested whether the guided Grad-CAM values at the maximum latency differed across brain regions. EEG channels were grouped into 6 brain regions (frontal cortex, central cortex, parietal cortex, occipital cortex, left temporal cortex, and right temporal cortex; Figure 4 – figure supplement 2). The electrodes for the 6 regions are as follows: Fp1, Fp2, Fz, F1, F2, AF3, AF4, AF7, AF8, and Fpz for the frontal region; C3, C4, Cz, FC1, FC2, C1, C2, FC3, and FC4 for the central region; P3, P4, Pz, CP1, CP2, P1, P2, CP3, CP4, P5, P6, and CPz for the parietal region; F3, F7, T7, P7, FC5, CP5, TP9, F5, C5, FT7, TP7, and FT9 for the left temporal region; and F4, F8, T8, P8, FC6, CP6, TP10, F6, C6, FT8, TP8, and FT10 for the right temporal region. The values obtained by each channel in a region were averaged, and regional differences were tested using Friedman’s test with the post-hoc Tukey–Kramer method ($P < 0.05$).

Data and code availability

All data supporting the findings of this study are provided within the paper. All custom Python codes used for deep learning and learned model data in this study are available at https://github.com/masaki-takeda/dld. All additional information will be made available upon reasonable request to the authors.

Acknowledgments

This study was supported by Kakenhi (Japan Society for the Promotion of Science) grant numbers 20H00521 and 21K18267 to MT; 21H00211 and 21K07262 to NW; 21H05060 to KJ; 21K20303 to DS; and 17H00891 to KN. This study was also supported by a grant from Uehara Memorial Foundation to MT, and a grant from Takeda Science Foundation to MT. We thank Dr. Shigeyuki Oba, Dr. Shinichi Yoshida, Mr. Zhen Zhang, and Dr. Makoto Iwata for their technical advice on deep learning. We also thank Mr. Rei Furutani and Ms. Maoko Yamanaka for their technical and administrative assistance, respectively.

Author contributions

N.W., K.J., K.N., and M.T. designed the experiment and study. N.W., R.K., and M.T. collected the data. N.W., K.M., and M.T. analyzed the data. N.W., K.M., K.J., D.S., K.N., and M.T. wrote the manuscript.

Competing interests

The authors declare no competing interests.

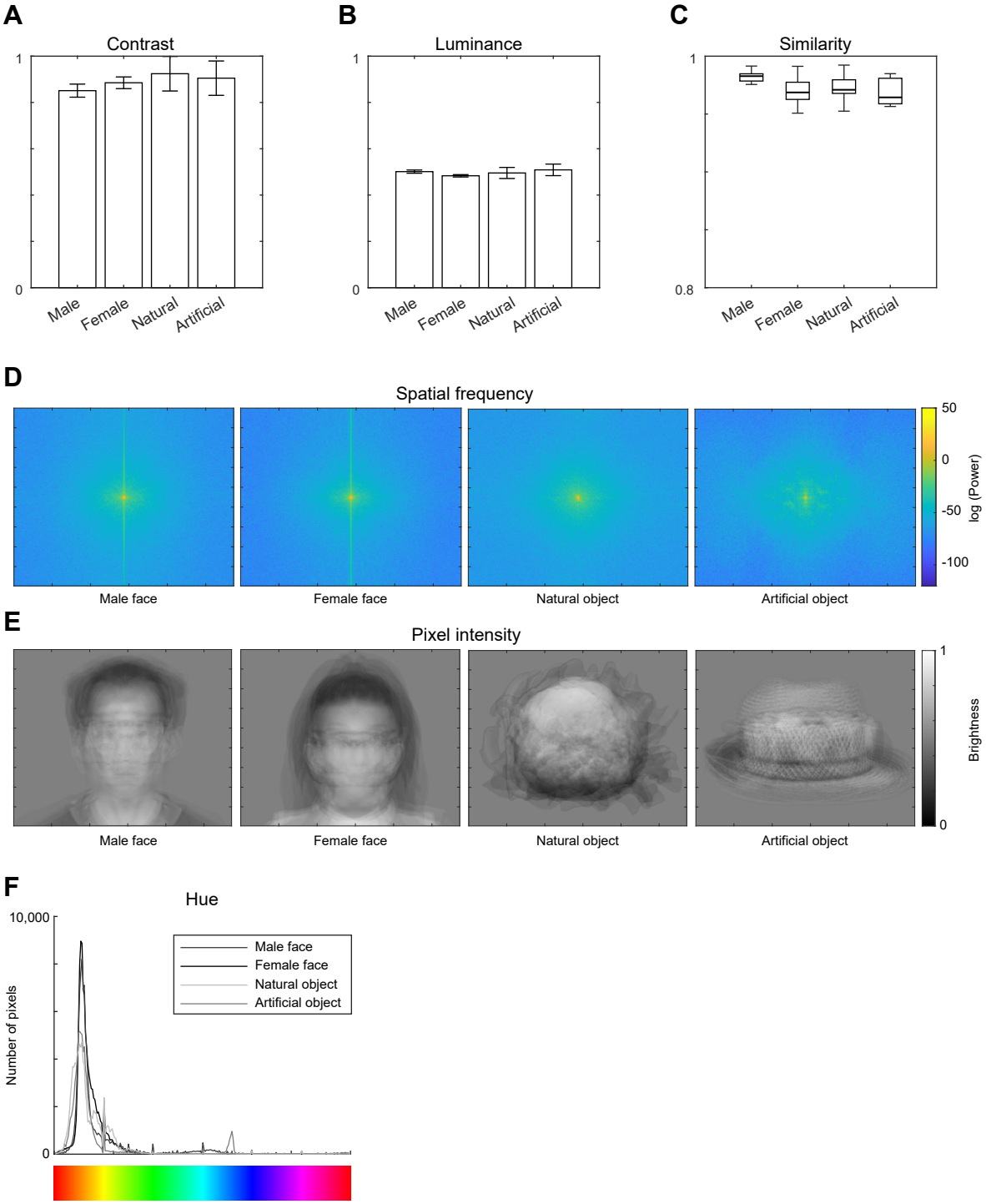

Figure 1 – figure supplement 1

Figure 1 – figure supplement 1. Low-level visual features of stimulus images. (A) Michelson contrast of images as the ratio between the difference and the sum of the maximum and minimum pixel intensities. **(B)** Luminance as the mean pixel intensity of images. **(C)** Pixel-based pairwise similarity. To estimate visual similarity among images of a category, the mean Euclidean distance between normalized grayscale values (scaled to be between 0 and 1) of each stimulus and every other stimulus of the same sub-category is measured. Pairwise similarity values range from 0 (images with inverted intensity difference at each pixel) to 1 (identical images). Note that the order of the size of the effect [0.0062, standard deviation (S.D.) of category means in normalized grayscale values] is comparable to the resolution of the display (0.0039, minimum pixel-wise distance resolved by the display in normalized values, corresponding to the change in one gray-level value of a possible 255 increments). Mean ± S.D. are shown in (A–C). **(D)** Average spatial frequency across images. **(E)** Average pixel intensity for each pixel. **(F)** Average hue histogram across images.

Figure 3 – figure supplement 1. Brain regions showing significant signal increase and decrease in the contrast of face vs. object. Coordinates are listed in MNI space.

Increase in face trial

Cluster index	Cluster name	t value	x (mm)	y (mm)	z (mm)	Harvard-Oxford Atlas
1	FMC	8.37	4	48	-10	Frontal Medial Cortex
		7.76	8	58	12	Frontal Pole
		6.87	-8	60	20	Frontal Pole
		5.69	-12	48	10	Paracingulate Gyrus
		4.89	4	32	6	Cingulate Gyrus, anterior division
		4.43	-24	42	-4	Frontal Pole
		4.17	-10	48	32	Frontal Pole
		4.04	-8	32	-6	Cingulate Gyrus, anterior division
		3.77	8	48	36	Superior Frontal Gyrus
2	R. AG	7.25	52	-56	18	Angular Gyrus
		6.76	40	-56	24	Angular Gyrus
		5.34	60	-38	48	Supramarginal Gyrus, posterior division
		5.34	54	-60	36	Lateral Occipital Cortex, superior division
		4.64	46	-42	10	Supramarginal Gyrus, posterior division
		4.48	64	-44	2	Middle Temporal Gyrus, temporooccipital part
		4.23	64	-46	34	Supramarginal Gyrus, posterior division
		3.67	46	-62	52	Lateral Occipital Cortex, superior division
3	PcC	6.49	6	-56	20	Precuneous Cortex
		6.28	-14	-52	32	Cingulate Gyrus, posterior division
		5.51	6	-56	34	Precuneous Cortex
		4.51	-12	-48	46	Precuneous Cortex
4	L. AG	5.30	-52	-52	20	Angular Gyrus
		5.27	-54	-64	38	Lateral Occipital Cortex, superior division
		4.77	-62	-52	34	Supramarginal Gyrus, posterior division
		4.59	-38	-60	18	Angular Gyrus
5	R. TOF	9.05	44	-48	-22	Temporal Occipital Fusiform Cortex
6	R. FOC	5.61	40	30	-12	Frontal Orbital Cortex
		4.44	26	14	-20	Frontal Orbital Cortex
7	L. MTGp	4.79	-66	-10	-24	Middle Temporal Gyrus, posterior division
8	R. LOCi	5.71	44	-80	-10	Lateral Occipital Cortex, inferior division
9	R. MTGp	5.37	60	-8	-16	Middle Temporal Gyrus, posterior division
		4.47	60	0	-26	Middle Temporal Gyrus, anterior division
10	R. Amygdala	8.46	20	-8	-16	Right Amygdala
11	L. LOCi	5.24	-44	-84	-6	Lateral Occipital Cortex, inferior division
		3.47	-50	-80	6	Lateral Occipital Cortex, inferior division
12	R. OcP	5.96	12	-98	12	Occipital Pole

Increase in object trial

Cluster index	Cluster name	t value	x (mm)	y (mm)	z (mm)	Harvard-Oxford Atlas
1	L. TOF/LOC	-15.40	-28	-52	-16	Temporal Occipital Fusiform Cortex
		-11.40	-50	-62	-8	Lateral Occipital Cortex, inferior division
		-10.60	-26	-70	-12	Occipital Fusiform Gyrus
		-9.75	-24	-64	46	Lateral Occipital Cortex, superior division
		-9.66	-30	-84	20	Lateral Occipital Cortex, superior division
		-9.64	-28	-86	4	Lateral Occipital Cortex, inferior division
		-9.44	-12	-94	-4	Occipital Pole
		-8.24	-26	-66	34	Lateral Occipital Cortex, superior division
		-7.55	-32	-30	-20	Temporal Fusiform Cortex, posterior division
		-7.31	-34	-50	44	Superior Parietal Lobule
		-6.69	-46	-30	40	Supramarginal Gyrus, anterior division
		-5.41	-12	-76	40	Precuneous Cortex
		-5.39	-24	-62	-28	Cerebellum
		-4.06	-50	-38	56	Postcentral Gyrus

		-3.86	-40	-64	-30	Cerebellum
2	R. TOF/LOC	-15.40	28	-46	-16	Temporal Occipital Fusiform Cortex
		-14.00	38	-86	14	Lateral Occipital Cortex, superior division
		-12.60	28	-62	-12	Occipital Fusiform Gyrus
		-10.10	50	-60	-8	Lateral Occipital Cortex, inferior division
		-9.81	24	-78	-10	Occipital Fusiform Gyrus
		-9.42	16	-88	0	Occipital Pole
		-7.47	28	-34	-22	Temporal Fusiform Cortex, posterior division
		-7.18	28	-68	50	Lateral Occipital Cortex, superior division
		-6.52	8	-76	-18	Lingual Gyrus
		-6.44	32	-82	32	Lateral Occipital Cortex, superior division
		-6.22	34	-72	-44	Cerebellum
		-5.98	30	-64	36	Lateral Occipital Cortex, superior division
		-5.81	14	-74	-44	Cerebellum
		-5.78	32	-68	24	Lateral Occipital Cortex, superior division
		-5.48	28	-62	-28	Cerebellum
		-5.44	22	-84	40	Lateral Occipital Cortex, superior division
		-5.10	10	-70	50	Precuneous Cortex
		-5.07	10	-84	-28	Occipital Fusiform Gyrus
		-4.88	32	-56	62	Superior Parietal Lobule
		-4.76	-8	-70	-42	Cerebellum
		-4.45	18	-64	60	Lateral Occipital Cortex, superior division
		-4.44	-26	-64	-48	Cerebellum
		-4.35	20	-60	-48	Cerebellum
		-4.14	0	-74	-32	Cerebellum
		-4.09	50	-44	-10	Inferior Temporal Gyrus, temporooccipital part
		-3.92	14	-56	-34	Cerebellum
		-3.84	-38	-56	-48	Cerebellum
		-3.45	-30	-60	-36	Cerebellum
		-3.44	26	-50	-34	Cerebellum
3	L. MFGi	-9.85	-48	10	34	Middle Frontal Gyrus
		-8.34	-46	32	14	Inferior Frontal Gyrus, pars triangularis
		-7.15	-42	24	24	Middle Frontal Gyrus
		-7.04	-30	24	4	Insular Cortex
		-4.85	-52	14	-6	Inferior Frontal Gyrus, pars opercularis
		-4.84	-30	32	-10	Frontal Orbital Cortex
		-3.98	-42	8	4	Central Opercular Cortex
		-3.92	-38	2	42	Precentral Gyrus
4	PcG	-6.81	6	10	50	Paracingulate Gyrus
		-6.63	-6	18	42	Paracingulate Gyrus
		-6.11	10	20	32	Cingulate Gyrus, anterior division
		-5.32	-8	4	54	Supplementary Motor Cortex
		-4.79	-10	-4	70	Superior Frontal Gyrus
		-4.66	-2	-10	58	Supplementary Motor Cortex
		-3.84	10	4	70	Superior Frontal Gyrus
5	L. MFGs	-5.60	-38	-2	60	Middle Frontal Gyrus
		-4.76	-22	2	58	Superior Frontal Gyrus
6	R. Ins	-5.60	32	24	4	Insular Cortex
		-4.57	54	16	-2	Inferior Frontal Gyrus, pars opercularis
		-3.79	52	14	-16	Temporal Pole
7	L. Thalamus	-6.39	-10	-18	10	Left Thalamus
		-4.05	-14	-4	16	Left Caudate
		-3.60	-10	-20	-8	Left Thalamus
8	R. SFG	-5.55	26	4	62	Superior Frontal Gyrus
		-3.96	36	-2	58	Middle Frontal Gyrus

Abbreviations. R, right; L, left, AG, angular gyrus; FG, fusiform gyrus; FMC, frontal medial cortex; FOC, frontal orbital cortex; Ins, insular cortex; LOC, lateral occipital cortex; MFG, middle frontal gyrus; MTG, middle temporal gyrus; OcP, occipital pole; OFC, orbitofrontal cortex; OP, occipital pole; PcC, precuneous cortex; PcG, paracingulate gyrus; SFG, superior frontal gyrus; TOF, temporal occipital fusiform cortex; i, inferior; s, superior; p, posterior

Figure 3 – figure supplement 2. Brain regions showing significant signal increase and decrease in the contrast of male-face vs. female-face. Coordinates are listed in MNI space.

Increase in male-face trial

Cluster index	t value	x (mm)	y (mm)	z (mm)	Harvard-Oxford Atlas

Increase in female-face trial

Cluster index	t value	x (mm)	y (mm)	z (mm)	Harvard-Oxford Atlas
1	-5.34	-30	-68	-12	Occipital Fusiform Gyrus
	-4.91	-16	-90	-6	Occipital Pole
	-4.87	-36	-78	-16	Occipital Fusiform Gyrus
	-4.84	-28	-72	22	Lateral Occipital Cortex, superior division
	-4.67	-20	-80	-14	Occipital Fusiform Gyrus
	-4.42	-42	-80	12	Lateral Occipital Cortex, inferior division
	-4.40	-54	-68	-8	Lateral Occipital Cortex, inferior division
	-4.31	-30	-54	-18	Temporal Occipital Fusiform Cortex
	-4.13	-34	-76	0	Lateral Occipital Cortex, inferior division
	-4.07	-16	-96	8	Occipital Pole
	-4.05	-48	-56	-16	Inferior Temporal Gyrus, temporooccipital part
	-4.02	-38	-88	0	Lateral Occipital Cortex, inferior division
	-3.61	-46	-78	-4	Lateral Occipital Cortex, inferior division
2	-4.84	28	-66	-14	Occipital Fusiform Gyrus
	-4.49	36	-46	-16	Temporal Occipital Fusiform Cortex
	-3.94	36	-88	-12	Lateral Occipital Cortex, inferior division
	-3.91	50	-70	-14	Lateral Occipital Cortex, inferior division
	-3.89	24	-82	-6	Occipital Fusiform Gyrus
	-3.67	42	-74	-2	Lateral Occipital Cortex, inferior division
3	-4.62	36	-76	4	Lateral Occipital Cortex, inferior division

Figure 3 – figure supplement 3. Brain regions showing significant signal increase and decrease in the contrast of natural-object vs. artificial-object. Coordinates are listed in MNI space.

Increase in natural-object trial

Cluster index	t value	x (mm)	y (mm)	z (mm)	Harvard-Oxford Atlas

Increase in artificial-object trial

Cluster index	t value	x (mm)	y (mm)	z (mm)	Harvard-Oxford Atlas
1	-5.61	38	-74	-6	Lateral Occipital Cortex, inferior division
	-5.45	38	-80	6	Lateral Occipital Cortex, inferior division
	-5.28	28	-78	-12	Occipital Fusiform Gyrus
	-4.75	34	-82	18	Lateral Occipital Cortex, superior division
	-3.95	42	-68	-16	Occipital Fusiform Gyrus
	-3.53	26	-78	0	Occipital Fusiform Gyrus
2	-4.18	-32	-92	12	Occipital Pole
	-3.66	-18	-94	22	Occipital Pole

Figure 3 – figure supplement 4. Guided Grad-CAM values for face and object trials in anatomically-defined brain regions. Values are sorted by their magnitudes. An asterisk denotes FDR q < 0.05.

Face trial

	Label	Region	Value	FDR *q*			Label	Region	Value	FDR *q*	
1	L. LOCi	Occipital	0.0067	1.10E-113	*	56	Brain-Stem	Subcortex	5.33E-06	1.06E-30	*
2	R. LOCi	Occipital	0.0061	1.16E-114	*	57	R. IFGpo	Frontal	4.42E-06	1.72E-05	*
3	R. TOF	Temporal	0.0023	3.85E-93	*	58	R. PhGa	Temporal	4.30E-06	0.0018	*
4	L. OcP	Occipital	0.0013	3.43E-113	*	59	R. SPL	Parietal	4.03E-06	2.99E-13	*
5	R. OcP	Occipital	0.00049	2.02E-112	*	60	R Accumbens	Subcortex	3.77E-06	1.53E-05	*
6	R. MTGtp	Temporal	0.00041	3.33E-89	*	61	L. PT	Temporal	3.65E-06	0.016	*
7	L. TOF	Temporal	0.00031	9.49E-58	*	62	L. ITGa	Temporal	3.45E-06	0.00014	*
8	L. IcC	Occipital	0.00012	1.22E-40	*	63	L. IFGpo	Frontal	3.01E-06	0.099	
9	R. AG	Parietal	0.00010	4.93E-33	*	64	L Thalamus	Subcortex	2.43E-06	5.62E-07	*
10	R. SmGp	Parietal	0.00010	2.74E-68	*	65	R Pallidum	Subcortex	2.23E-06	0.079	
11	L. COpc	Frontal	8.85E-05	8.35E-99	*	66	R. MFG	Frontal	2.03E-06	0.0056	*
12	L. PrG	Frontal	7.39E-05	1.11E-108	*	67	L. MTGa	Temporal	2.03E-06	0.23	
13	L. LG	Occipital	7.27E-05	7.71E-34	*	68	R Caudate	Subcortex	1.93E-06	0.0025	*
14	R. LG	Occipital	7.26E-05	2.95E-31	*	69	L Caudate	Subcortex	1.91E-06	4.15E-05	*
15	L. SMC	Frontal	5.43E-05	6.79E-98	*	70	R. ITGa	Temporal	1.62E-06	0.017	*
16	L. PoG	Parietal	5.40E-05	2.75E-103	*	71	L. STGa	Temporal	1.22E-06	0.022	*
17	L. H1/H2	Parietal	4.98E-05	3.52E-39	*	72	R. SFG	Frontal	1.21E-06	0.0041	*
18	L. LOCs	Parietal	4.92E-05	6.98E-51	*	73	R. PcC	Parietal	6.89E-07	0.0017	*
19	R. LOCs	Parietal	4.40E-05	7.36E-47	*	74	R. TP	Temporal	6.14E-07	0.54	
20	L. POpc	Parietal	3.88E-05	1.12E-37	*	75	R. STGp	Temporal	5.48E-07	0.54	
21	L. PcC	Parietal	3.84E-05	2.00E-67	*	76	L. FMC	Frontal	1.49E-07	0.28	
22	L. SmGa	Parietal	3.82E-05	1.66E-36	*	77	R Thalamus	Subcortex	1.25E-07	0.31	
23	R. CGa	Frontal	3.65E-05	2.41E-60	*	78	R. SmGa	Parietal	-8.87E-07	0.59	
24	L. SPL	Parietal	2.94E-05	5.83E-85	*	79	L. PhGa	Temporal	-1.01E-06	0.54	
25	L. CGa	Frontal	2.90E-05	7.12E-57	*	80	L Accumbens	Subcortex	-1.27E-06	1	
26	L. Ins	Insula	2.84E-05	1.07E-28	*	81	R. Ins	Insula	-1.96E-06	0.91	
27	L. SmGp	Parietal	2.27E-05	5.32E-13	*	82	R. MTGa	Temporal	-2.18E-06	1	
28	R. ScC	Frontal	2.24E-05	1.56E-53	*	83	L. MFG	Frontal	-2.39E-06	1	
29	L. PP	Temporal	2.21E-05	3.47E-19	*	84	L. ScC	Frontal	-2.39E-06	1	
30	L Putamen	Subcortex	2.21E-05	1.64E-41	*	85	R. STGa	Temporal	-2.79E-06	1	
31	R. SccC	Occipital	2.00E-05	0.00018	*	86	L. SFG	Frontal	-3.02E-06	1	
32	R. FMC	Frontal	1.95E-05	2.15E-37	*	87	R. PP	Temporal	-4.09E-06	1	
33	L. OFG	Occipital	1.85E-05	1		88	L. STGp	Temporal	-6.18E-06	1	
34	R Amygdala	Subcortex	1.78E-05	1.78E-28	*	89	L Pallidum	Subcortex	-7.76E-06	1	
35	R. SMC	Frontal	1.76E-05	1.82E-47	*	90	R. FP	Frontal	-9.19E-06	1	
36	R. CGp	Parietal	1.75E-05	9.07E-36	*	91	R. MTGp	Temporal	-9.29E-06	1	
37	L. TFCp	Temporal	1.71E-05	6.56E-11	*	92	L. CGp	Parietal	-9.78E-06	1	
38	R. TFCa	Temporal	1.66E-05	1.10E-40	*	93	L. FP	Frontal	-1.02E-05	1	
39	L. FOpc	Frontal	1.63E-05	8.63E-29	*	94	L. IFGpt	Frontal	-1.02E-05	1	
40	R. IcC	Occipital	1.60E-05	0.0034	*	95	L. PhGp	Temporal	-1.16E-05	1	
41	L. TP	Temporal	1.60E-05	2.96E-46	*	96	L. MTGp	Temporal	-1.30E-05	1	
42	R Putamen	Subcortex	1.53E-05	3.82E-27	*	97	L. ITGp	Temporal	-1.54E-05	1	
43	R. FOC	Frontal	1.49E-05	9.81E-49	*	98	R. H1/H2	Parietal	-1.65E-05	1	
44	L. PcG	Frontal	1.46E-05	7.25E-21	*	99	L Hippocampus	Subcortex	-2.84E-05	1	
45	R. PcG	Frontal	1.46E-05	8.66E-14	*	100	R. POpC	Parietal	-6.59E-05	1	
46	R. COpC	Frontal	1.39E-05	3.19E-13	*	101	R. ITGp	Temporal	-7.27E-05	1	
47	L. AG	Parietal	1.16E-05	0.0043	*	102	R Hippocampus	Subcortex	-7.60E-05	1	
48	R. TFCp	Temporal	1.11E-05	0.042	*	103	L. CC	Occipital	-8.47E-05	1	
49	R. PoG	Parietal	1.01E-05	1.73E-41	*	104	R. CC	Occipital	-9.99E-05	1	
50	R. PrG	Frontal	9.93E-06	2.87E-39	*	105	R. PhGp	Temporal	-0.00012	1	
51	R. IFGpt	Frontal	9.56E-06	4.09E-20	*	106	L. MTGtp	Temporal	-0.00014	1	
52	L. TFCa	Temporal	6.93E-06	9.98E-07	*	107	R. PT	Temporal	-0.00016	1	
53	L. FOC	Frontal	5.89E-06	2.01E-19	*	108	R. ITGtp	Temporal	-0.00033	1	
54	R. FOpC	Frontal	5.63E-06	9.07E-09	*	109	L. ITGtp	Temporal	-0.00039	1	
55	L Amygdala	Subcortex	5.61E-06	1.46E-06	*	110	R. OFG	Occipital	-0.00040	1	

Object trial

	Label	Region	Value	FDR *q*			Label	Region	Value	FDR *q*	
1	L. TOF	Temporal	0.0046	7.44E-92	*	56	R. SmGa	Parietal	8.67E-07	0.27	
2	L. ITGtp	Temporal	0.0035	9.32E-93	*	57	L. FP	Frontal	6.47E-07	0.0019	*
3	R. TOF	Temporal	0.0014	5.46E-84	*	58	L. ScC	Frontal	3.93E-07	0.12	
4	L. LOCi	Occipital	0.0014	6.64E-86	*	59	L. PoG	Parietal	3.44E-07	0.00012	*
5	L. OcP	Occipital	0.00088	8.90E-95	*	60	L Accumbens	Subcortex	2.39E-07	0.012	*
6	R. ITGtp	Temporal	0.00084	2.63E-76	*	61	L. PhGa	Temporal	2.26E-07	0.68	
7	R. OFG	Occipital	0.00066	6.97E-87	*	62	R. Ins	Insula	3.67E-08	1	
8	R. OcP	Occipital	0.00060	8.90E-95	*	63	L Putamen	Subcortex	-1.04E-08	1	
9	L. OFG	Occipital	0.00049	2.21E-71	*	64	R Accumbens	Subcortex	-2.52E-08	0.73	
10	R. LOCi	Occipital	0.00045	1.47E-68	*	65	R. TP	Temporal	-3.08E-07	0.52	
11	R. LOCs	Parietal	0.00020	9.14E-84	*	66	R Pallidum	Subcortex	-4.39E-07	1	
12	L. ITGp	Temporal	0.00010	1.70E-16	*	67	L Pallidum	Subcortex	-4.97E-07	1	
13	L. LOCs	Parietal	9.80E-05	3.04E-60	*	68	L. TP	Temporal	-1.10E-06	1	
14	L. TFCp	Temporal	7.93E-05	6.54E-18	*	69	R. COpC	Frontal	-1.24E-06	1	
15	R. FOpC	Frontal	5.52E-05	4.75E-55	*	70	L. IFGpo	Frontal	-1.28E-06	0.30	
16	R. SPL	Parietal	3.93E-05	1.22E-67	*	71	R. MFG	Frontal	-1.43E-06	1	
17	L. IFGpt	Frontal	3.74E-05	1.15E-18	*	72	L. PcG	Frontal	-1.70E-06	1	
18	R. IcC	Occipital	3.49E-05	7.43E-12	*	73	L. FOpc	Frontal	-2.24E-06	0.62	
19	L. SPL	Parietal	2.82E-05	7.56E-47	*	74	R. SFG	Frontal	-2.71E-06	1	
20	L. PrG	Frontal	2.71E-05	9.67E-62	*	75	R. H1/H2	Parietal	-2.91E-06	1	
21	L. LG	Occipital	2.66E-05	5.88E-06	*	76	L. STGa	Temporal	-2.97E-06	1	
22	R. TFCp	Temporal	2.62E-05	9.81E-11	*	77	R. PP	Temporal	-3.36E-06	1	
23	R Caudate	Subcortex	1.83E-05	2.79E-24	*	78	R Amygdala	Subcortex	-3.41E-06	1	
24	L. SMC	Frontal	1.76E-05	3.32E-09	*	79	L Amygdala	Subcortex	-3.51E-06	1	
25	R. PcC	Parietal	1.73E-05	1.89E-26	*	80	R. ScC	Frontal	-3.55E-06	1	
26	L. POpc	Parietal	1.50E-05	1.16E-29	*	81	L. Ins	Insula	-3.58E-06	1	
27	R. CC	Occipital	1.30E-05	2.72E-05	*	82	L. FMC	Frontal	-3.67E-06	1	
28	L. H1/H2	Parietal	1.30E-05	3.78E-14	*	83	L. SmGa	Parietal	-3.78E-06	1	
29	R. ITGa	Temporal	1.05E-05	2.85E-33	*	84	L. SFG	Frontal	-3.92E-06	1	
30	R Thalamus	Subcortex	9.74E-06	1.80E-34	*	85	L. COpc	Frontal	-4.58E-06	1	
31	R. CGa	Frontal	8.03E-06	1.18E-17	*	86	R. FP	Frontal	-5.28E-06	1	
32	L. ITGa	Temporal	7.80E-06	1.20E-29	*	87	R. FMC	Frontal	-5.34E-06	1	
33	R. CGp	Parietal	6.71E-06	0.00098	*	88	R. SccC	Occipital	-5.57E-06	1	
34	R. PrG	Frontal	6.39E-06	2.71E-21	*	89	R. TFCa	Temporal	-7.03E-06	1	
35	R. PoG	Parietal	5.97E-06	1.22E-38	*	90	R. AG	Parietal	-8.02E-06	1	
36	R. SMC	Frontal	5.66E-06	2.33E-10	*	91	R. STGp	Temporal	-8.02E-06	1	
37	R. PcG	Frontal	5.60E-06	3.42E-09	*	92	L. SmGp	Parietal	-1.04E-05	1	
38	R. LG	Occipital	5.42E-06	1		93	L. CC	Occipital	-1.05E-05	1	
39	L. FOC	Frontal	4.93E-06	1.92E-17	*	94	R. SmGp	Parietal	-1.19E-05	1	
40	L. TFCa	Temporal	4.65E-06	5.25E-11	*	95	R. ITGp	Temporal	-1.26E-05	1	
41	L Caudate	Subcortex	4.58E-06	1.58E-13	*	96	L. STGp	Temporal	-1.37E-05	1	
42	R. IFGpo	Frontal	4.52E-06	4.07E-10	*	97	R. MTGtp	Temporal	-1.55E-05	1	
43	Brain-Stem	Subcortex	3.49E-06	2.52E-15	*	98	L. PP	Temporal	-1.59E-05	1	
44	R. STGa	Temporal	3.12E-06	3.21E-08	*	99	L. IcC	Occipital	-1.80E-05	1	
45	L. CGp	Parietal	3.01E-06	0.0040	*	100	R. IFGpt	Frontal	-1.88E-05	1	
46	L Thalamus	Subcortex	2.81E-06	5.80E-08	*	101	R. MTGp	Temporal	-1.91E-05	1	
47	L. CGa	Frontal	2.70E-06	0.0051	*	102	L Hippocampus	Subcortex	-1.91E-05	1	
48	R. POpC	Parietal	2.68E-06	3.36E-05	*	103	L. PcC	Parietal	-2.22E-05	1	
49	L. MTGa	Temporal	2.68E-06	0.0031	*	104	L. PT	Temporal	-3.06E-05	1	
50	R. MTGa	Temporal	2.63E-06	3.37E-07	*	105	R Hippocampus	Subcortex	-4.91E-05	1	
51	R. FOC	Frontal	2.07E-06	1.96E-06	*	106	L. AG	Parietal	-5.36E-05	1	
52	R. PhGa	Temporal	1.69E-06	0.0082	*	107	R. PhGp	Temporal	-5.48E-05	1	
53	L. MFG	Frontal	1.62E-06	0.0067	*	108	L. MTGp	Temporal	-6.69E-05	1	
54	R Putamen	Subcortex	1.30E-06	0.042	*	109	L. PhGp	Temporal	-0.00011	1	
55	R. PT	Temporal	9.14E-07	0.17		110	L. MTGtp	Temporal	-0.00030	1	

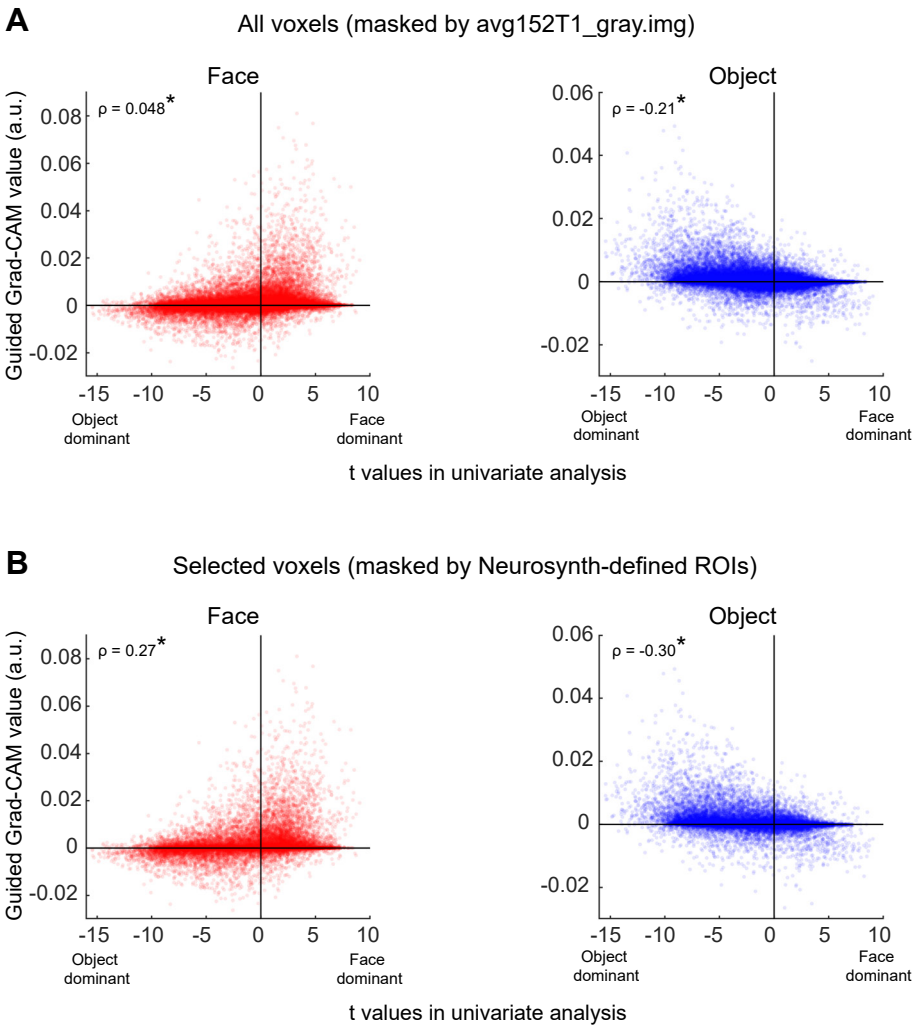

Figure 3 – figure supplement 5

Figure 3 – figure supplement 5. Voxel-based relationship between t-values in the univariate analysis and the guided Grad-CAM values in the face and object trial. (A) Relationship between the two values when voxels are masked by FSL's gray matter prior volume. **(B)** Relationship between the two values when voxels are masked by functional ROIs obtained from Neurosynth using the terms "face" and "object." *, $P < 0.001$.

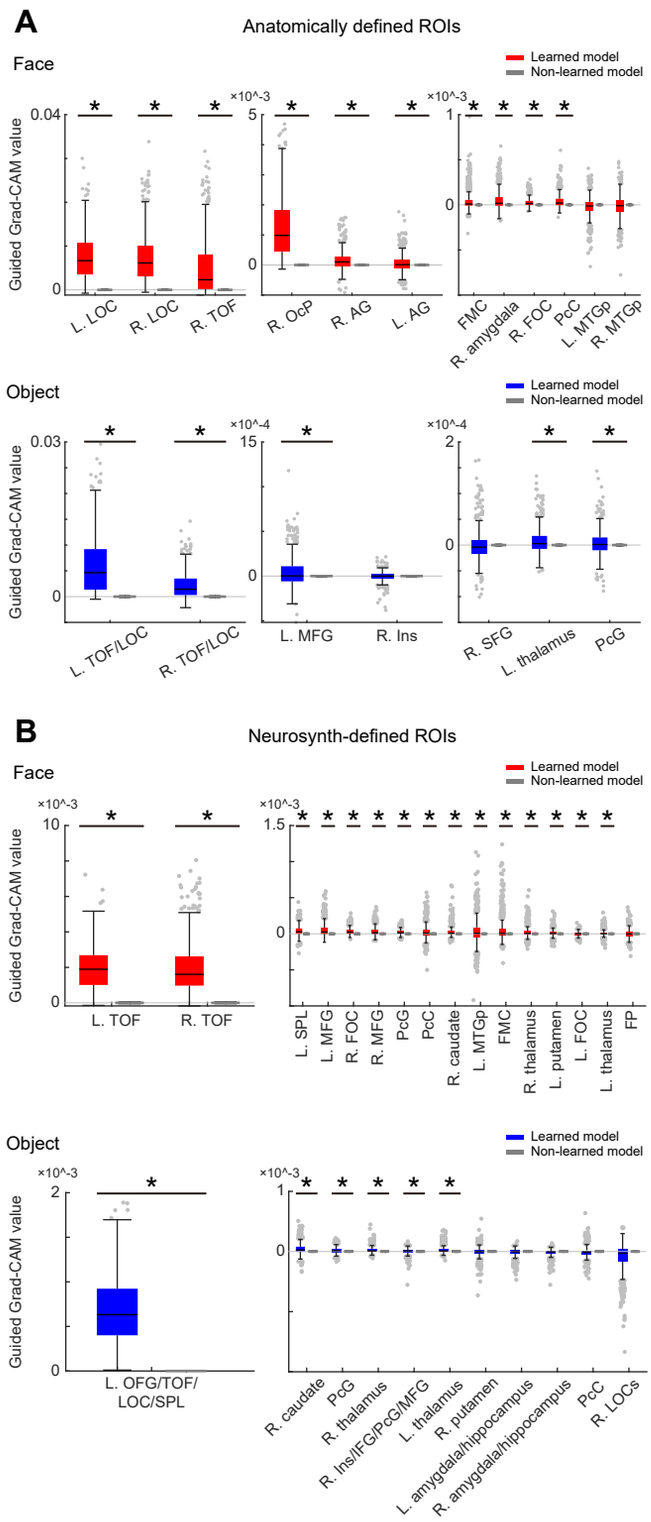

Figure 3 – figure supplement 6

Figure 3 — figure supplement 6. ROI analysis. Guided Grad-CAM values for each ROI when ROIs are defined **(A)** anatomically (Harvard–Oxford atlas) and **(B)** functionally (Neurosynth). Figure configurations are the same as in Figure 3f. L, left; R, right; AG, angular gyrus; FMC, frontal medial cortex; FOC, frontal orbital cortex; FP, frontal pole; IFG, inferior frontal gyrus; Ins, insular cortex; LOC, lateral occipital cortex; MFG, middle frontal gyrus; MTG, middle temporal gyrus; OcP, occipital pole; OFG, occipital fusiform gyrus; PcC, precuneous cortex; PcG, paracingulate gyrus; SFG, superior frontal gyrus; SPL, superior parietal lobule; TOF, temporal occipital fusiform cortex; i, inferior; s, superior; p, posterior.

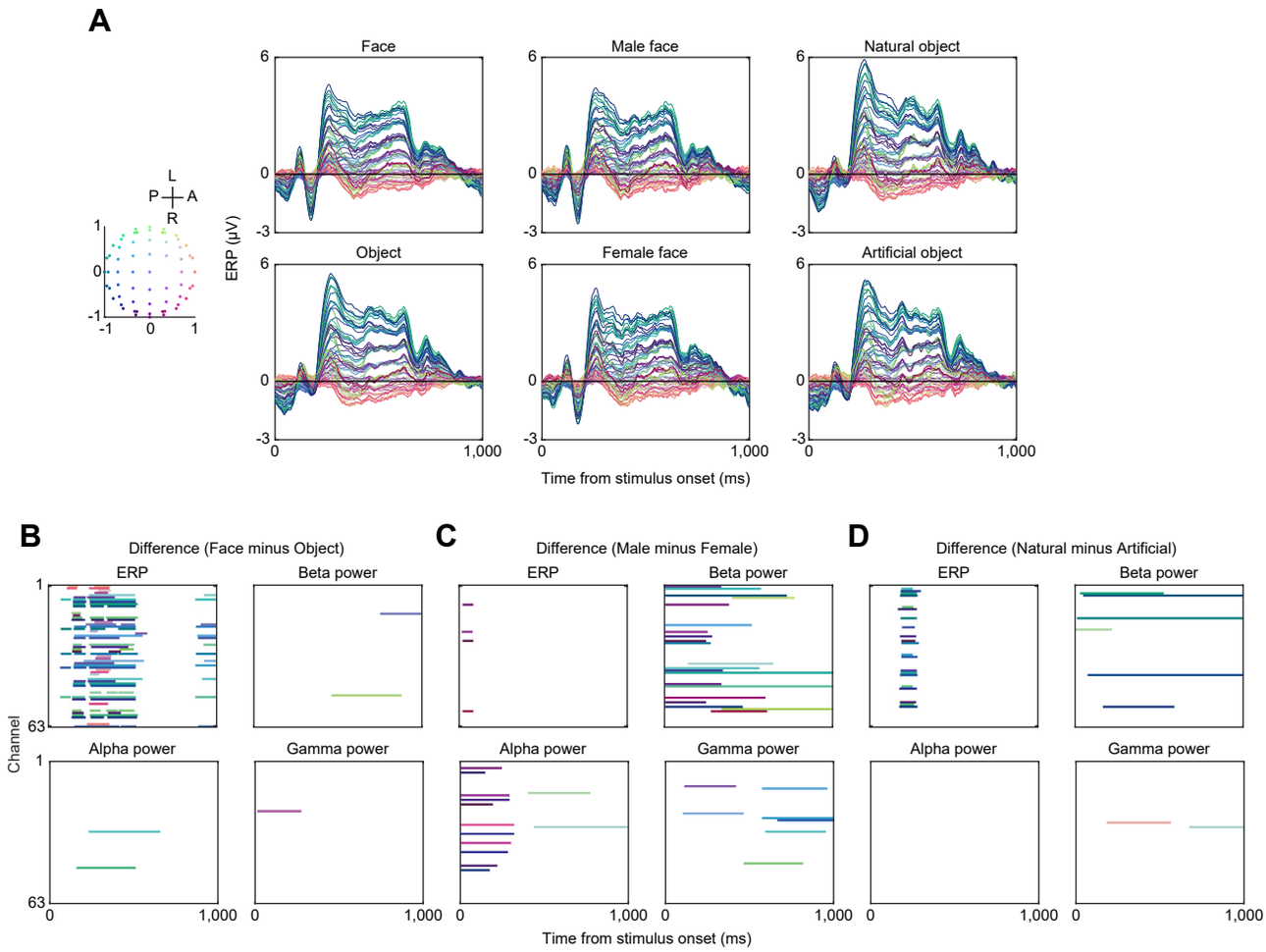

Figure 4 – figure supplement 1

Figure 4 — figure supplement 1. EEG responses. (A) ERPs for face trials, object trials, male face trials, female face trials, natural object trials, and artificial object trials. Line colors correspond to the channel location depicted on the left side. A, anterior, P, posterior, R, right, L, left. **(B–D)** Significance of temporal trajectories in ERPs and power spectrograms. Colored lines depict time points showing significant differential ERP amplitude or power spectrogram against zero in respective channels. The color of each line depicts the location of the corresponding channels (left). Time points that exceeded the 97.5^{th} percentile of the permutation distribution (N = 5,000) served as cluster-inducing time points, and the significance of clusters is set at the threshold of $P < 0.05$.

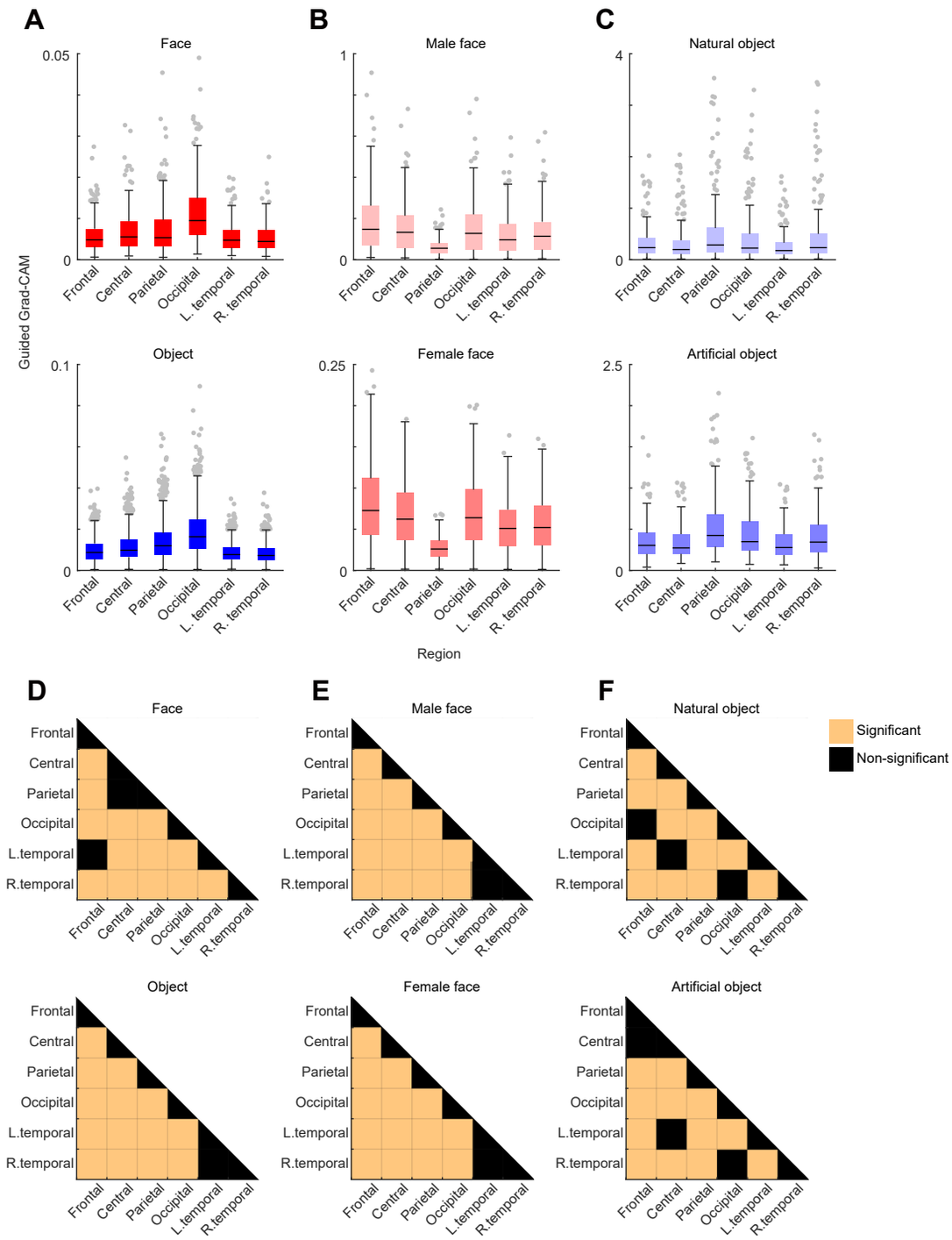

Figure 4 – figure supplement 2

Figure 4 — figure supplement 2. Regional analysis of guided Grad-CAM. (A–C) Comparison of the guided Grad-CAM values at the maximum latency in face and object trials (A), in male face and female face trials (B), and in natural object and artificial object trials (C) across six regions. All 63 channels are categorized as either frontal, central, parietal, occipital, left temporal, and right temporal regions based on their locations. Shown are trial-based median value distributions across channels in a region. Regional differences in the maximum guided Grad-CAM values are statistically significant in all trial types (Friedman's test, $P < 0.001$). **(D–F)** Results of multiple comparison test (Tukey–Kramer method, $P < 0.05$). Significant differences in the guided Grad-CAM values between regions are color-coded.

References

Allen, P.J., Josephs, O., and Turner, R. (2000). A method for removing imaging artifact from continuous EEG recorded during functional MRI. Neuroimage *12*, 230-239. 10.1006/nimg.2000.0599.

Bai, S., Kolter, J.Z., and Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networksfor Sequence Modeling. arXiv:1803.01271v2. 10.48550/arXiv.1803.01271.

Bar, M. (2004). Visual objects in context. Nat Rev Neurosci *5*, 617-629. 10.1038/nrn1476.

Barton, J.J. (2011). Disorder of higher visual function. Curr Opin Neurol *24*, 1-5. 10.1097/WCO.0b013e328341a5c2.

Birn, R.M., Diamond, J.B., Smith, M.A., and Bandettini, P.A. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. Neuroimage *31*, 1536-1548. 10.1016/j.neuroimage.2006.02.048.

Carlson, T., Tovar, D.A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. J Vis *13*, 1. 10.1167/13.10.1.

Cavina-Pratesi, C., Kentridge, R.W., Heywood, C.A., and Milner, A.D. (2010). Separate processing of texture and form in the ventral stream: evidence from FMRI and visual agnosia. Cereb Cortex *20*, 433-446. 10.1093/cercor/bhp111.

Cichy, R.M., and Oliva, A. (2020). A M/EEG-fMRI Fusion Primer: Resolving Human Brain Responses in Space and Time. Neuron *107*, 772-781. 10.1016/j.neuron.2020.07.001.

Cichy, R.M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. Nat Neurosci *17*, 455-462. 10.1038/nn.3635.

Dobs, K., Isik, L., Pantazis, D., and Kanwisher, N. (2019). How face perception unfolds over time. Nat Commun *10*, 1258. 10.1038/s41467-019-09239-1.

Eklund, A., Nichols, T.E., and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A *113*, 7900-7905. 10.1073/pnas.1602413113.

Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature *392*, 598-601. 10.1038/33402.

Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science *291*, 312-316. 10.1126/science.291.5502.312.

Frey, M., Nau, M., and Doeller, C.F. (2021). Magnetic resonance-based eye tracking using deep neural networks. Nat Neurosci *24*, 1772-1779. 10.1038/s41593-021-00947-w.

Gainotti, G., and Marra, C. (2011). Differential contribution of right and left temporo-occipital and anterior temporal lesions to face recognition disorders. Front Hum Neurosci *5*, 55. 10.3389/fnhum.2011.00055.

Geusebroek, J., Burghouts, G.J., and Smeulders, A.W.M. (2005). The Amsterdam Library of Object Images. Int. J. Comput. Vis. *61*, 103-112. 10.1023/B:VISI.0000042993.50813.60.

Grill-Spector, K., and Weiner, K.S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. Nat Rev Neurosci *15*, 536-548. 10.1038/nrn3747.

Grill-Spector, K., Weiner, K.S., Kay, K., and Gomez, J. (2017). The Functional Neuroanatomy of Human Face Perception. Annu Rev Vis Sci *3*, 167-196. 10.1146/annurev-vision-102016-061214.

Harvey, A.K., Pattinson, K.T., Brooks, J.C., Mayhew, S.D., Jenkinson, M., and Wise, R.G. (2008). Brainstem functional magnetic resonance imaging: disentangling signal from physiological noise. J Magn Reson Imaging *28*, 1337-1344. 10.1002/jmri.21623.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science *293*, 2425-2430. 10.1126/science.1063736.

Hong, H., Yamins, D.L., Majaj, N.J., and DiCarlo, J.J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. Nat Neurosci *19*, 613-622. 10.1038/nn.4247.

Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. Nat Commun *8*, 15037. 10.1038/ncomms15037.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863-866. 10.1126/science.1117593.

Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., and Haxby, J.V. (1999). Distributed representation of objects in the human ventral visual pathway. Proc Natl Acad Sci U S A *96*, 9379-9384. 10.1073/pnas.96.16.9379.

Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., and Wei, Y. (2021). LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. IEEE Trans Image Process *30*, 5875-5888. 10.1109/TIP.2021.3089943.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci *17*, 4302-4311. 10.1523/JNEUROSCI.17-11-04302.1997.

Kasper, L., Bollmann, S., Diaconescu, A.O., Hutton, C., Heinzle, J., Iglesias, S., Hauser, T.U., Sebold, M., Manjaly, Z.M., Pruessmann, K.P., and Stephan, K.E. (2017). The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. J Neurosci Methods *276*, 56-72. 10.1016/j.jneumeth.2016.10.019.

Keerativittayayut, R., Aoki, R., Sarabi, M.T., Jimura, K., and Nakahara, K. (2018). Large-scale network integration in the human brain tracks temporal fluctuations in memory encoding performance. Elife *7*, e32696. 10.7554/eLife.32696.

Kingma, D., and Ba, J. (2015). Adam: A Method for Stochastic Optimization. ICLR 2015. 10.48550/arXiv.1412.6980.

Konen, C.S., Behrmann, M., Nishimura, M., and Kastner, S. (2011). The functional neuroanatomy of object agnosia: a case study. Neuron *71*, 49-60. 10.1016/j.neuron.2011.05.030.

Kourtzi, Z., and Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. J Neurosci *20*, 3310-3318. 10.1523/JNEUROSCI.20-09-03310.2000.

Kreiman, G., Koch, C., and Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. Nat Neurosci *3*, 946-953. 10.1038/78868.

Kriegeskorte, N., Formisano, E., Sorger, B., and Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci U S A *104*, 20600-20605. 10.1073/pnas.0705654104.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126-1141. 10.1016/j.neuron.2008.10.043.

Landi, S.M., Viswanathan, P., Serene, S., and Freiwald, W.A. (2021). A fast link between face perception and memory in the temporal pole. Science *373*, 581-585. 10.1126/science.abi6671.

Liu, J., Harris, A., and Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. Nat Neurosci *5*, 910-916. 10.1038/nn909.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Annu Rev Neurosci *19*, 577-621. 10.1146/annurev.ne.19.030196.003045.

Mack, M.L., and Palmeri, T.J. (2011). The timing of visual object categorization. Front Psychol *2*, 165. 10.3389/fpsyg.2011.00165.

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc Natl Acad Sci U S A *92*, 8135-8139. 10.1073/pnas.92.18.8135.

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods *164*, 177-190. 10.1016/j.jneumeth.2007.03.024.

Matsui, T., Hattori, Y., Tsumura, K., Aoki, R., Takeda, M., Nakahara, K., and Jimura, K. (2022). Executive control by fronto-parietal activity explains counterintuitive decision behavior in complex value-based decision-making. Neuroimage *249*, 118892. 10.1016/j.neuroimage.2022.118892.

Miyakawa, N., Majima, K., Sawahata, H., Kawasaki, K., Matsuo, T., Kotake, N., Suzuki, T., Kamitani, Y., and Hasegawa, I. (2018). Heterogeneous Redistribution of Facial Subcategory Information Within and Outside the Face-Selective Domain in Primate Inferior Temporal Cortex. Cereb Cortex *28*, 1416-1431. 10.1093/cercor/bhx342.

Miyashita, Y., and Chang, H.S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature *331*, 68-70. 10.1038/331068a0.

Nestor, A., Plaut, D.C., and Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proc Natl Acad Sci U S A *108*, 9998-10003. 10.1073/pnas.1102433108.

Peelen, M.V., and Downing, P.E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. Neuropsychologia *105*, 177-183. 10.1016/j.neuropsychologia.2017.03.033.

Raschka, S. (2020). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv:1811.12808v3. 10.48550/arXiv.1811.12808.

Rokach, L. (2009). Ensemble-based classifiers. Artificial Intelligence Review *33*, 1-39. 10.1007/s10462-009-9124-7.

Rousselet, G.A., Fabre-Thorpe, M., and Thorpe, S.J. (2002). Parallel processing in high-level categorization of natural images. Nat Neurosci *5*, 629-630. 10.1038/nn866.

Sarraf, S., DeSouza, D.D., Anderson, J., and Tofighi, G. (2017). DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. bioRxiv. 10.1101/070441.

Schiltz, C., Sorger, B., Caldara, R., Ahmed, F., Mayer, E., Goebel, R., and Rossion, B. (2006). Impaired face discrimination in acquired prosopagnosia is associated with abnormal response to individual faces in the right middle fusiform gyrus. Cereb Cortex *16*, 574-586. 10.1093/cercor/bhj005.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision *128*, 336-359. 10.1007/s11263-019-01228-7.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci U S A *104*, 6424-6429. 10.1073/pnas.0700622104.

Simonyan, K., and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR2015. 10.48550/arXiv.1409.1556.

Stigliani, A., Weiner, K.S., and Grill-Spector, K. (2015). Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. J Neurosci *35*, 12412-12424. 10.1523/JNEUROSCI.4822-14.2015.

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. IEEE Trans Neural Syst Rehabil Eng *25*, 1998-2008. 10.1109/TNSRE.2017.2721116.

Suzuki, W.A., and Naya, Y. (2014). The perirhinal cortex. Annu Rev Neurosci *37*, 39-53. 10.1146/annurev-neuro-071013-014207.

Takeda, M., Hirabayashi, T., Adachi, Y., and Miyashita, Y. (2018). Dynamic laminar rerouting of inter-areal mnemonic signal by cognitive operations in primate temporal cortex. Nat Commun *9*, 4629. 10.1038/s41467-018-07007-1.

Tanaka, D., Aoki, R., Suzuki, S., Takeda, M., Nakahara, K., and Jimura, K. (2020). Self-Controlled Choice Arises from Dynamic Prefrontal Signals That Enable Future Anticipation. J Neurosci *40*, 9736-9750. 10.1523/JNEUROSCI.1702-20.2020.

Tanaka, K. (1996). Inferotemporal cortex and object vision. Annu Rev Neurosci *19*, 109-139. 10.1146/annurev.ne.19.030196.000545.

Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B., and Tootell, R.B. (2003). Faces and objects in macaque cerebral cortex. Nat Neurosci *6*, 989-995. 10.1038/nn1111.

Tsao, D.Y., Freiwald, W.A., Tootell, R.B., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670-674. 10.1126/science.1119983.

Tsumura, K., Aoki, R., Takeda, M., Nakahara, K., and Jimura, K. (2021a). Cross-Hemispheric Complementary Prefrontal Mechanisms during Task Switching under Perceptual Uncertainty. J Neurosci *41*, 2197-2213. 10.1523/JNEUROSCI.2096-20.2021.

Tsumura, K., Kosugi, K., Hattori, Y., Aoki, R., Takeda, M., Chikazoe, J., Nakahara, K., and Jimura, K. (2021b). Reversible Fronto-occipitotemporal Signaling Complements Task Encoding and Switching under Ambiguous Cues. Cereb Cortex, bhab324. 10.1093/cercor/bhab324.

Turesson, H.K., Logothetis, N.K., and Hoffman, K.L. (2012). Category-selective phase coding in the superior temporal sulcus. Proc Natl Acad Sci U S A *109*, 19438-19443. 10.1073/pnas.1217012109.

Vu, H., Kim, H.C., Jung, M., and Lee, J.H. (2020). fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations. Neuroimage *223*, 117328. 10.1016/j.neuroimage.2020.117328.

Wang, X., Liang, X., Jiang, Z., Nguchu, B.A., Zhou, Y., Wang, Y., Wang, H., Li, Y., Zhu, Y., Wu, F., et al. (2020). Decoding and mapping task states of the human brain via deep learning. Hum Brain Mapp *41*, 1505-1519. 10.1002/hbm.24891.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G.O., Gosselin, F., and Tanaka, J.W. (2010). Controlling low-level image properties: the SHINE toolbox. Behav Res Methods *42*, 671-684. 10.3758/BRM.42.3.671.

Worsley, K.J., and Friston, K.J. (1995). Analysis of fMRI time-series revisited--again. Neuroimage *2*, 173-181. 10.1006/nimg.1995.1023.

Yamins, D.L., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci *19*, 356-365. 10.1038/nn.4244.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nat Methods *8*, 665-670. 10.1038/nmeth.1635.

Yu, F., Koltun, V., and Funkhouser, T. (2017). Dilated Residual Networks. CVPR 2017, 636-644. 10.1109/cvpr.2017.75.

Zhang, Z., Yu, X., Rong, X., and Iwata, M. (2021). Spatial-Temporal Neural Network for P300 Detection. IEEE Access *9*, 163441-163455. 10.1109/access.2021.3132024.

Zhao, Y., Dong, Q., Zhang, S., Zhang, W., Chen, H., Jiang, X., Guo, L., Hu, X., Han, J., and Liu, T. (2018). Automatic Recognition of fMRI-Derived Functional Networks Using 3-D Convolutional Neural Networks. IEEE Trans Biomed Eng *65*, 1975-1984. 10.1109/TBME.2017.2715281.

Zheng, X., Mondloch, C.J., and Segalowitz, S.J. (2012). The timing of individual face recognition in the brain. Neuropsychologia *50*, 1451-1461. 10.1016/j.neuropsychologia.2012.02.030.