Received: 24 February 2022 | Revised: 4 December 2022 | Accepted: 12 December 2022

DOI: 10.1002/hbm.26189

RESEARCH ARTICLE

WILEY

Neural representations of the perception of handwritten digits and visual objects from a convolutional neural network compared to humans

Juhyeon Lee | Minyoung Jung | Niv Lustig | Jong-Hwan Lee

Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

Correspondence
Jong-Hwan Lee, Department of Brain and Cognitive Engineering, Korea University, Anam-ro 145, Seongbuk-gu, Seoul 02841, Republic of Korea.
Email: jonghwan_lee@korea.ac.kr

Funding information
Electronics and Telecommunications Research Institute, Grant/Award Number: 22ZS1100; National Research Foundation of Korea, Grant/Award Numbers: 2017R1E1A1A01077288, 2021M3E5D2A01022515

Abstract

We investigated neural representations for visual perception of 10 handwritten digits and six visual objects from a convolutional neural network (CNN) and humans using functional magnetic resonance imaging (fMRI). Once our CNN model was fine-tuned using a pre-trained VGG16 model to recognize the visual stimuli from the digit and object categories, representational similarity analysis (RSA) was conducted using neural activations from fMRI and feature representations from the CNN model across all 16 classes. The encoded neural representation of the CNN model exhibited the hierarchical topography mapping of the human visual system. The feature representations in the lower convolutional (Conv) layers showed greater similarity with the neural representations in the early visual areas and parietal cortices, including the posterior cingulate cortex. The feature representations in the higher Conv layers were encoded in the higher-order visual areas, including the ventral/medial/dorsal stream and middle temporal complex. The neural representations in the classification layers were observed mainly in the ventral stream visual cortex (including the inferior temporal cortex), superior parietal cortex, and prefrontal cortex. There was a surprising similarity between the neural representations from the CNN model and the neural representations for human visual perception in the context of the perception of digits versus objects, particularly in the primary visual and associated areas. This study also illustrates the uniqueness of human visual perception. Unlike the CNN model, the neural representation of digits and objects for humans is more widely distributed across the whole brain, including the frontal and temporal areas.

KEYWORDS
convolutional neural network, functional magnetic resonance imaging, handwritten digits, representational similarity analysis, visual objects, visual perception

Juhyeon Lee and Minyoung Jung contributed equally.

1 | INTRODUCTION

Recently, a number of studies have reported the similarity between the primate visual pathway and convolutional neural network (CNN)-based models, which are the best-performing computational models for visual object recognition (Bracci et al., 2019; Cichy et al., 2016; Cohen et al., 2020; Eickenberg et al., 2017; Güçlü & van Gerven, 2015; King et al., 2019; Lindh et al., 2019; Mehrer et al., 2021; Wen et al., 2018). For example, Güçlü and van Gerven systematically evaluated the similarity of the feature representations in each CNN layer with the neural activations indirectly measured from functional magnetic resonance imaging (fMRI) along the visual pathway, including the lateral occipital complex. The lower and higher layers of the CNN resembled the topographical organization of the human visual areas (Güçlü & van Gerven, 2015). In another study, CNN-based computational models showed remarkable similarity with inferior temporal (IT) cortex representation, outperforming bio-inspired object recognition models such as HMAX (Cadieu et al., 2014).

Representational similarity analysis (RSA), which can be used to evaluate relationships across various modalities such as brain activity data, computational models, and behavioral data, has been instrumental in previous brain-encoding studies (Kriegeskorte et al., 2008). The construction of a representational dissimilarity matrix (RDM), which consists of dissimilarity scores for the modalities between pairs of experimental stimuli/conditions, is a crucial component of RSA that is used to relate heterogeneous modalities in holistic geometric space across stimuli/conditions (Kriegeskorte et al., 2008). CNN-based encoding models for visual perception have been successfully employed in decoding models for visual stimuli reconstruction (Shen et al., 2019; Wen et al., 2018). More recently, a brain-encoding model based on a CNN was used to develop an ecological dataset that could maximize the similarity between the trained CNN model and the visual object perception of the human brain (Mehrer et al., 2021). However, despite the large number of previous brain-encoding studies, very few have investigated the neural representations of symbolic visual stimuli using human brain activation and computer vision models.

Numeral digits are fundamental symbolic visual stimuli for humans that encode the numerical concept of abstract numbers (Dehaene, 1992; Nieder, 2016). The neural underpinnings of numerical digit recognition in the human brain have long been the focus of research (Anobile et al., 2021; Ansari et al., 2007; Bulthé et al., 2014; Dehaene, 1992; Dehaene & Cohen, 1995; Nieder, 2021; Yeo et al., 2020). Bulthé and colleagues demonstrated that multivoxel pattern analysis (MVPA) of fMRI data can be used to decode the numerical magnitude of Arabic digits from the localized regions. More recently, Yeo and colleagues reported the categorical distinction of numbers versus other symbols in the number-preferring region in the posterior inferior temporal gyrus (pITG). However, very few studies have investigated the visual perception of symbolic digits in conjunction with that of objects. Investigating neural representations for visual object recognition simultaneously with symbolic digit recognition is fundamental to understanding human visual perception. We believe that brain-encoding research on the visual perception of concrete objects and symbolic digits is urgently needed.

In the present study, we were motivated to investigate the neural representation of digit and object perception from a CNN model and humans using naturalistic visual stimuli. To this end, we used the ten handwritten digits available in the MNIST dataset and six objects in the ImageNet dataset as visual stimuli. The neural activations for each of the stimuli were acquired from fMRI data. The features from the CNN model trained to classify the 10 digits and six objects were represented as an RDM across all 16 classes. Consequently, we conducted RSA to obtain the neural representation of the human brain from a trained CNN model perspective. We also designed RDM codes that encode several hypothetical human perceptions across our digit and object categories in terms of the numerical magnitude of the digits (i.e., small vs. large) and animacy of the objects (i.e., animate vs. inanimate) in addition to distinguishing between digits, between objects, and digits versus objects conditions. We believed that the systematic comparison of the neural representations from the CNN model and from human perception would illustrate their similarities and differences, providing an insight into the development of CNN-based computational models that imitate the visual perception of humans.

2 | MATERIALS AND METHODS

2.1 | Overview

Figure 1 summarizes our investigation of the neural representations of handwritten digit and visual object recognition using two RSA scenarios. First, we constructed a neural RDM for the 10 handwritten digits and six visual objects using the multivoxel patterns of neural activations in a searchlight area for each voxel using measured fMRI data (Figure 1a). To investigate the neural representations of visually presented category/class perception from the CNN, we fine-tuned a pre-trained VGG16-based CNN model for the classification of the 16 classes used in our study. Subsequently, we obtained an RDM for each of the CNN layers. To investigate the neural representations of visual perception from humans, an RDM that encoded visual category/class perception was designed (Figure 1b). Searchlight RSA was conducted between the neural RDM and the RDMs of (i) the CNN model (see Section 2.9 for more details) or (ii) human visual perception (see Section 2.10 for more details). The potential links between the neural representations obtained from the two RSA approaches were also examined across two stages. In the first RSA, we computed the similarity between the RDM for the CNN model and the RDM for the category/class perception of humans. In a subsequent RSA, we obtained the similarity between the RDMs for each CNN layer and the neural RDMs obtained from the representative regions-of-interest (ROIs) for the category/class perception of humans in the first RSA. Please see Section 2.10 for the ROI definition and Section 2.11 for the subsequent RSA descriptions.

2.2 | Participants

The Institutional Review Board (IRB) at Korea University approved the entire study protocol. All participants submitted written consent forms and were compensated as described in the IRB documents. Twenty-three healthy right-handed volunteers (mean ± standard deviation = 22.3 ± 2.6 years old; 10 females, 13 males; Edinburg's handedness score = 87.3 ± 10.4) with no self-reported neurological or neuropsychiatric problems participated in the study.

2.3 | MRI parameters

MRI data were acquired using a 3T Siemens MAGNETOME MRI scanner with a 20-channel head coil (Siemens, Erlangen, Germany). Blood-oxygen-ation-level-dependent (BOLD) fMRI data were acquired using a multi-band gradient-echo echo-planar-imaging (EPI) pulse sequence developed from the Center for Magnetic Resonance Research (CMRR, Department of Radiology, University of Minnesota). Specific imaging parameters were multiband factor = 2, in-plane multiband Generalized Auto-Calibrating Partial Parallel Acquisitions, time of repetition (TR) = 1440 ms, echo time (TE) = 30 ms, flip angle (FA) = 71°, field-of-view (FoV) = 192 × 192 mm^2, 3 mm isotropic voxels, 50 interleaved slices without a gap, and 442 measurements in one fMRI run. A T1-weighted structural MRI volume was acquired using a magnetization-prepared rapid gradient-echo pulse sequence (TR = 1900 ms, TE = 2.28 ms, FA = 8°, FoV = 256 × 256 mm^2, and 1 mm isotropic voxels).

2.4 | Visual stimuli for the fMRI experiment

In the present study, we used a subset of the fMRI data collected for our research project to understand the processing of image, sound, and multimodal images with sound stimuli by the human brain. We used (a) ImageNet (https://www.image-net.org/) for visual object images, (b) MNIST (http://yann.lecun.com/exdb/mnist/) for handwritten digit images, and (c) the Google Speech Commands dataset for sound stimuli for the images of the objects and digits (https://pyroomacoustics.readthedocs.io/en/pypirelease/pyroomacoustics.datasets.google_speech_commands.html). From these image and sound stimuli, 10 digits (0–9) and six objects (bed, bird, cat, dog, house, and tree) were available in both modalities. Thus, these 16 classes were used for both the image and sound stimuli. More specifically, images of the objects in the six classes from the validation data for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC; 2012 version) (Deng et al., 2009) and grayscale images of handwritten digits in the 10 classes in the test data from the MNIST (Lecun et al., 1998) were used as visual stimuli for the fMRI experiment. Please refer to the supplementary materials, "Image stimuli for the fMRI experiment for the digit and object categories" for more detail.

2.5 | Task paradigm for the fMRI experiment

We randomly selected 12 samples for (a) each of the 16 (i.e., 10 digits and six objects) image classes, (b) each of the 16 sound classes, and (c) each of the 16 multimodal classes with an image and sound for each participant using the NumPy's "random" module in Python 3.6 (i.e., 192 images, 192 sounds, and 192 images with sound). These 576 stimuli/trials were interspersed across three fMRI runs in a pseudo-randomized order across unimodality/multimodality conditions and the 16 classes. There were 64 counter-balanced trials for images only, sounds only, and images with sound in one fMRI run (Figure 2). An image stimulus was presented for 0.2 s three times interleaved with a brief fixation cross for 0.2 s (i.e., a stimulus duration of 1 s) followed by a baseline fixation cross for 2 s. A sound stimulus

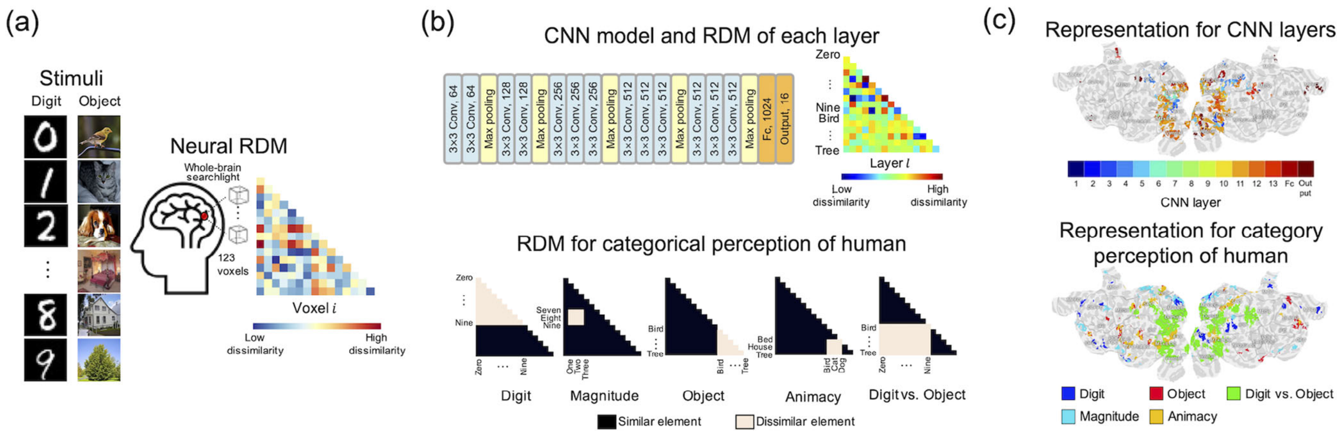

FIGURE 1 Visual perception of handwritten digits and visual objects under two representational similarity analysis (RSA) scenarios using human brain activations, a convolutional neural network (CNN), and human visual category/class perception. (a) Construction of a neural representational dissimilarity matrix (RDM) consisting of neural activations measured for the 10 handwritten digits and six objects images via fMRI. (b) Construction of (i) an RDM for each layer of the CNN trained to classify the 16 classes and (ii) an RDM that encodes the category/class visual perception of humans. (c) RSA using the neural RDM with (i) the RDM for the CNN model and (ii) the RDM for the category/class visual perception of humans. Fc, fully connected layer; Output, output layer.

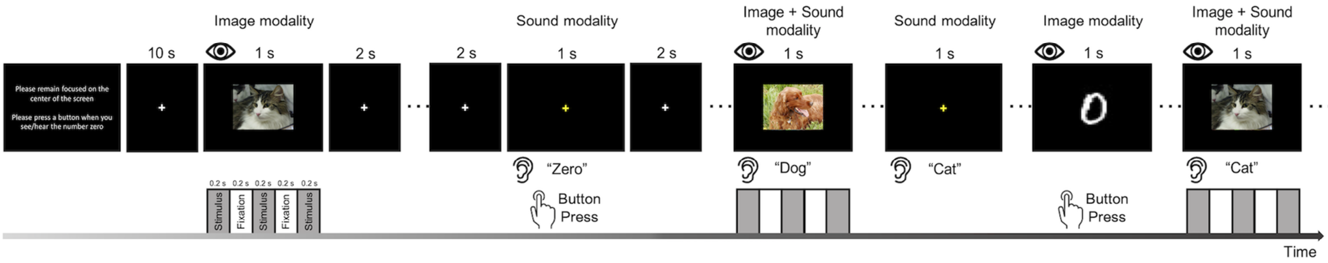

FIGURE 2 Experimental paradigm for fMRI data acquisition. Ten handwritten digits and six visual objects are presented as visual stimuli for the image modality trials. Sound waveforms corresponding to the visual stimuli are presented as sound stimuli for the sound modality trials. In the image + sound modality trials, images with corresponding sound stimuli are presented to participants as a multimodal condition. Participants are instructed to press a button whenever they see and/or hear the "0" digit to maintain their alertness throughout the experiment. Please refer to the Sections 2.4 and 2.5 for details

(1 s) was delivered to participants, followed by a 2 s fixation cross. The same stimulus duration and interval used for the unimodal conditions were used for the multimodal images with the sound. Participants were instructed to press an MR-compatible button response pad (Current Design, https://www.curdes.com) when the "0" image or sound stimulus was presented to maintain their level of alertness throughout the experiment.

A projector system (PROPixx; https://vpixx.com/products/propixx) with a mirror in the head coil and optional vision correction glasses (Mediglasses; https://www.crsltd.com) were used during the visual presentation stage. An MR-compatible active noise canceling auditory headset (OptoACTIVE Slim Optical ANC Headphones; https://www.optoacoustics.com) was used to deliver the sound. The MATLAB environment (version R2017b) was used to execute the experimental paradigm during fMRI acquisition and record the participants' responses. Three participants were excluded due to technical glitches in the stimulus presentation computer and the button response pad. Five participants were excluded due to poor task performance (>3 missed "0" stimuli from a total of 36 stimuli), mainly due to drowsiness. As a result, we used the fMRI data acquired from the remaining 15 participants.

2.6 | fMRI preprocessing

We employed a standard preprocessing pipeline with the "afni_proc.py" command from Analysis of Functional Neuroimages (AFNI) software (http://afni.nimh.nih.gov/afni) consisting of despiking, slice timing correction, realignment, the co-registration of fMRI volumes to an individual anatomical volume, spatial normalization to Montreal Neurological Institute (MNI) space, spatial smoothing using a 3D Gaussian kernel with an 8 mm full-width at half-maximum (FWHM), and scaling to set the mean BOLD intensity of each voxel to 100. The preprocessed fMRI data were analyzed using a general linear model (GLM) at an individual level. There were 192 regressors for each of the 192 task-related trials in one fMRI run. Six head motion parameters obtained from the preprocessing step with their derivatives, temporal drift artifacts modeled using up to five polynomial orders, and three principal components (PCs) extracted from the WM and CSF based on aCompCor (Behzadi et al., 2007) were also added as nuisance regressors (23 in total) to the GLM to account for non-neural artifacts in the BOLD signals. Neural activations from each trial were estimated from the beta values of the GLM across the whole brain. As a result, there were 192 beta-valued whole-brain maps available for the images only, sounds only, and multimodal images with sound for each participant. In this study, we used 192 beta-valued maps obtained from the image-only condition (i.e., 12 images for each of the 16 classes) per participant to investigate the neural representations of handwritten digits and visual objects.

2.7 | Training of the CNN model for the recognition of handwritten digits and visual objects

We used a pretrained VGG16-based CNN model to set the initial weights for our CNN model with 13 convolutional (Conv) layers (https://www.robots.ox.ac.uk/~vgg/research/very_deep/) (Simonyan & Zisserman, 2015). The final Conv layer of our CNN model was connected to a fully connected (Fc) layer with 1024 nodes followed by 16 output nodes in the output layer to recognize the 16 classes across the 10 digits and six objects. We used the rectified linear unit as an activation function for all hidden nodes. Cross-entropy between a target one-hot vector and predicted output vector in the output layer was used as the cost function to fine-tune the parameters of our CNN model. The TensorFlow library (version 1.15.0) was used to implement the CNN model. An adaptive moment estimation (Adam) optimizer was used (Kingma & Ba, 2017). The learning rate was initialized at 10^{-5} and exponentially decreased by 0.96 every 10 epochs. We applied dropout with a probability of 0.5 to the output of the Fc layer. Alternative to VGG16, we additionally trained AlexNet (Krizhevsky et al., 2017) and ResNet-50 (He et al., 2016). Please refer to the Section "Generalization on alternative CNN models" in Supplementary Materials for details.

We used MNIST ($n = 60,000$ images) and ImageNet ($n = 234,329$ images) data from the original training dataset for the training of our CNN model (Table S1). Each image was resampled to 224 × 224 pixels, and the grayscale MNIST image was copied to each of the three RGB channels to match the dimensions of the CNN input for the color image. To reduce the degree of potential overfitting, we used a data augmentation scheme to increase the number of training samples per class by applying varying degrees of shift and scaling to the original training images. More specifically, cropping, horizontal mirroring, and color jittering (Krizhevsky et al., 2017) were used for the object images (Simonyan & Zisserman, 2015) and rotation (between −15° and 15°), translation (between −8% and 8%), and scaling (between factors of 1 and 1.08) were used for the digit images (Shorten & Khoshgoftaar, 2019). Once the data augmentation schemes were applied, the difference in the number of training samples across the categories and classes (i.e., class imbalance problem) was mitigated by applying sub-sampling of the number of images for each class at each epoch in the training phase. Specifically, the number of training samples for each category was same as 96,660 (i.e., 16,110 for each of the six object classes, and 9666 for each of the 10 digit classes) in each epoch.

2.8 | Evaluation of the trained CNN model

We evaluated the performance of our trained CNN model using the validation data provided by ImageNet and MNIST (Table S1). We also assessed our trained CNN model in terms of (a) representative input patterns that maximized each of the hidden nodes and output nodes and (b) feature representation in each of the layers. First, the representative input pattern for each hidden node was estimated using the activation maximization (AM) method (Erhan et al., 2009). Specifically, the input pattern that maximized the activation of the target hidden node was estimated from a random noise pattern by applying a gradient ascent scheme. Several techniques were used to enhance the corresponding feature representations, such as Gaussian blurring using DeepDraw (https://github.com/auduno/deepdraw), which is based on the DeepDream code by Google Research (https://github.com/google/deepdream). High-dimensional feature representations in each layer were also interpreted in the 2D plane using *t*-distributed stochastic neighbor embedding (*t*-SNE) conducted via Barnes-Hut approximation (Van der Maaten & Hinton, 2008).

2.9 | RSA of visual category/class perception from the CNN

We conducted whole-brain searchlight RSA using the neural RDM and the RDM of feature representations for each of the CNN layers to examine the neural representations from the CNN model. In detail, the 12 beta-valued GLM maps obtained from each of the 12 image stimuli per class at an individual level were averaged. A multivoxel pattern of neural activations from the center voxel in a searchlight area was then obtained in a spherical region with a three-voxel radius (123 voxels), including the center voxel. Consequently, we constructed a neural RDM for the center voxel using the 16 vectors of 123 × 1 multivoxel patterns across the 16 classes, in which the Pearson's correlation coefficient (CC) *r* was used for a dissimilarity measure (i.e., 1 – *r*; minimum of 0 and a maximum of 2).

We obtained the RDM for each layer of the CNN using the corresponding features of the input image. The dimensions of the feature vector that was 1D concatenated across all features obtained from all Conv filters for one input image in each of the CNN layers was very large (i.e., from 100,352 to 3,211,264). Thus, we applied a dimension reduction method using principal component analysis (PCA). The number of PCs that preserved 90% of the variance of all feature representations from the 192 images across all 16 classes were obtained for each layer. Table S2 summarizes the dimensions of the 1D concatenated features in each CNN layer and the minimum number of PCs that preserved at least 90% of the explained variance. We then averaged the 12 sets of dimension-reduced features for 12 images per class in each CNN layer. Consequently, an RDM of feature representations across the 16 classes was constructed for each CNN layer using the dissimilarity measure based on Pearson's CCs (i.e., 1 – *r*) calculated across all pairs of the average dimension-reduced feature vectors from the 16 classes.

We conducted searchlight RSA using (a) the RDM for each layer of our CNN model and (b) the neural RDM obtained from each center voxel by moving the center voxel across the whole brain. To this end, Spearman's rank correlation (ρ) was used as a similarity measure between the neural RDM for each voxel and the RDM for each of the CNN layers. Fisher's *z*-transform was applied to the ρ values across the whole brain for each subject. A voxel-wise one-sample *t*-test was employed using the *z*-transformed ρ values across the 15 subjects for group inference. Multiple comparison correction of the resulting *p*-value was applied using random permutations ($n = 5000$). More specifically, multivoxel patterns with randomized voxel indices were used to construct a neural RDM using an individual beta-valued map followed by RSA with the RDM for the CNN layers based on Spearman's rank correlation. The resulting ρ values across the 15 subjects were Fisher's *z*-transformed and subject to the one-sample *t*-test. Thus, we obtained a null distribution of *t*-statistics to correct the *p*-value obtained from the RSA using intact voxel indices in the multivoxel pattern. As a result, the clusters that exhibited significant similarity (corrected $p < 0.05$) between the neural RDM and the RDM for each CNN layer were determined from cluster-size correction with a minimum of 15 voxels.

We identified the label of the CNN layers that showed the largest *t*-scores across all CNN layers for each of the voxels in the significant clusters as a summary map of the RSA. We referred to this map across the whole brain as the CNN layer assignment map and visualized it as a flat cortical map using Pycortex (Gao et al., 2015) (https://github.com/gallantlab/pycortex). We employed independent component analysis (ICA) to parcellate the whole brain area based on the functional information as measured from a BOLD time series (Smith et al., 2009), in which 100 independent components were estimated. We labeled the functionally parcellated regions based on Glasser's 360 multimodal parcellations (Glasser et al., 2016) and the Human Brainnetome Atlas (Fan et al., 2016). We visualized the clusters and the labeled functional boundaries on the cortical surface using Pycortex. Fine-grained information on the strength of the association between the CNN layer and the neural activations was analyzed using bar graphs across the CNN layers for each of the significant clusters in the CNN layer assignment map. We also performed the RSA using individual beta-valued maps obtained from preprocessed fMRI data without spatial smoothing. We conducted the group inference using the resulting individual RSA maps via the one-sample *t*-test. Alternatively, the individual RSA maps were spatially smoothed with 4 mm FWHM before the *t*-test. Please see the supplementary materials, "RSA using unsmoothed beta-valued brain maps" for the details.

2.10 | RSA of visual category/class perception from humans

We also conducted whole-brain searchlight RSA using RDMs that could potentially encode human visual category/class perception across the ten handwritten digits and six visual objects. To this end, we prepared five sets of 16 × 16 RDMs as shown in Figure 1b that encoded (i) handwritten digit perception (i.e., ones in the RDM elements across the ten digits, otherwise zeros), (ii) magnitude information for the digits (i.e., lower [1, 2, 3] vs. higher [7, 8, 9]), (iii) visual object perception (ones in the RDM elements across the six objects, otherwise zeros), (iv) animacy information for the objects (i.e., animate [bird, cat, and dog] vs. inanimate [bed, house, and tree]), and (v) the perception of the digits in comparison to the objects (ones in the RDM elements between the digits and objects, otherwise zeros). Consequently, the RSA was conducted using the neural RDM for each voxel and the RDM codes to identify the ROIs encoding each concept in holistic geometric space across 16 classes. As with the searchlight RSA for the CNN model, we calculated the voxel-wise Spearman's rank correlation (ρ) between the neural RDM and the RDM codes at an individual level. One-sample *t*-tests were then used with the Fisher's *z*-transformed ρ values from all of the 15 participants for group inference. Multiple comparison correction of the *p*-values was employed using random permutations ($n = 10{,}000$) followed by a cluster-level threshold with a minimum of 20 connected voxels (using "3dClusterize" in AFNI with the "edges touching" option and with a corrected $p < 0.05$). The resulting voxel clusters were defined as ROIs. We also obtained the individual RSA maps from unsmoothed beta-values followed by group inference via the one-sample *t*-test using (a) individual RSA maps and (b) spatially smoothed individual RSA maps with 4 mm FWHM. Please see the supplementary materials, "RSA using unsmoothed beta-valued brain maps" for the details.

Although we assigned [1,2,3] and [7,8,9] for the grouping of lower and higher digits, the grouping cannot be uniquely determined. Thus, we also constructed the RDMs using other groups of the lower and higher digits from the digit "5" by excluding the digit "0." Now, there were four digits in each of the two groups (i.e., [1,2,3,4] for lower and [6,7,8,9] for higher). Following that, the magnitude RDMs were constructed based on all the 16 combinatorial conditions (i.e., three digits selected for each of the lower and higher magnitude groups), and a subsequent RSA was performed. The resulting 16 RSA maps in each subject were averaged and one-sample *t*-test was performed using the average RSA maps across the 15 participants, followed by multiple comparison correction. In addition, we also investigated the neural representations of the digit versus object condition using an alternative set of six object classes to evaluate the generalizability of our findings. Please refer to the supplementary materials, "RSA using an alternative set of six objects" for the detailed methods and results.

2.11 | Links between the trained CNN model and the visual category/class perception of humans

We examined the potential links between the two RSA perspectives obtained from the CNN model and the visual perception of humans. First, we evaluated the degree of visual perception of humans in the trained CNN model using the cosine similarity between the RDM for each of the CNN layers and the RDM for each visual category/class perception of humans (Figure 1b). We also conducted RSA between the neural RDMs in the ROIs found from the category/class perception codes of humans and the RDM for each layer of the CNN model. Only the dissimilar elements in the RDM (bottom of Figure 1b) that encodes the category/class perception of interest (i.e., across digits, magnitude of digits, across objects, animacy of objects, and digits vs. objects) were used for this RSA. The similarity between the neural RDM in the ROIs and the RDM for each CNN layer was calculated using Spearman's rank correlation ρ. One-sample *t*-tests were applied to *z*-transformed ρ values across 15 participants for group inference. We averaged positive *t*-scores within each ROI to reveal the relationship between the ROIs representing the visual perception of humans and each of the CNN layers.

3 | RESULTS

3.1 | Evaluation of the trained CNN model

The average training accuracy and average validation accuracy across the 16 classes were 99.8% and 99.2%, respectively. Figure S1 shows the confusion matrix of classification for the validation data. Figure 3a illustrates the estimated input patterns that maximized each of the three randomly selected nodes for each layer of the trained CNN model, which illustrates the hierarchical processing of the visual information across the layers. Figure 3b shows the estimated input patterns from the AM of each output node that visualized each of the 16 classes. Figure 4 visualizes the *t*-SNE plots of the feature representations in the CNN layers, including the input and output layers. The *t*-SNE plots clearly show that specific information regarding classes was obtained from the Fc and output layers rather than from the Conv layers. Based on the degree of separation across the 16 classes in the *t*-SNE representations, we divided the CNN Conv layers into lower (Conv 1 and Conv 2), intermediate (Conv 3–10), and higher

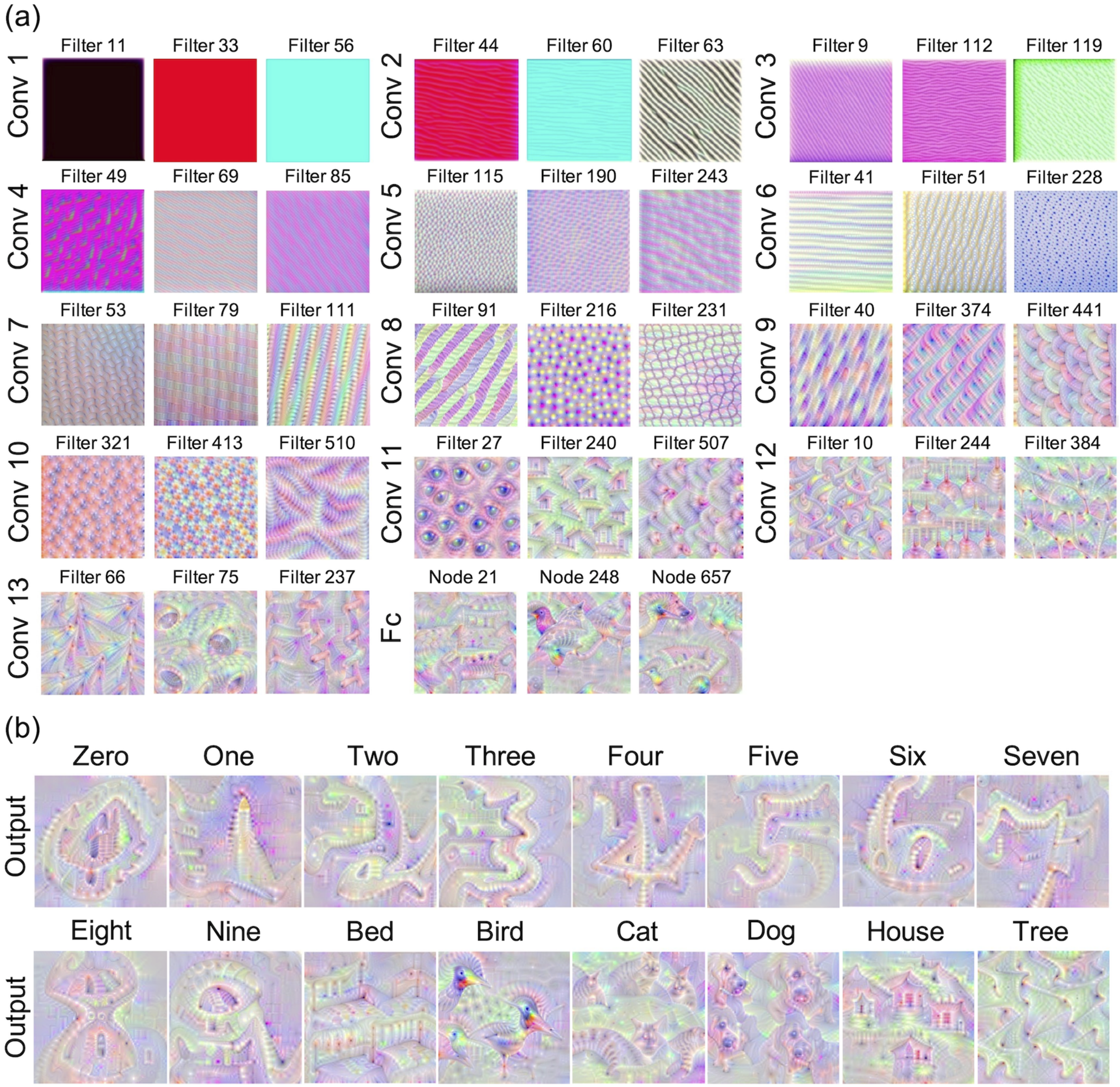

FIGURE 3 Evaluation of the trained CNN model. (a) Estimated input patterns are obtained using the activation maximization (AM) approach, in which the estimated input pattern indicates the most representative patterns for the corresponding convolutional layer (Conv) filters or nodes of the fully connected (Fc) layer. For the visualization, Conv filters and nodes at the Fc layer are randomly selected. (b) Estimated input patterns from the AM applied to each of the 16 output nodes. Output, output layer.

(Conv 11–13) layers. We defined the Fc and output layers of the CNN as the classification layers.

3.2 | Neural representations from the perspective of the trained CNN model

Figure 5 shows the CNN layer assignment map across the whole brain and bar graphs denoting the similarity between the neural activations and feature representations in each CNN layer at the group level. Overall, it is worth noting that the maximum association between the CNN layer and human brain activations showed a hierarchy within the visual areas and across the whole brain from the visual to the frontal areas (Figure 5a). In the lower Conv layers, the feature representations exhibited their maximum association with the neural activations in the right somatomotor areas (M1/S1) and the right middle temporal (MT) complex and its neighboring visual areas, including the lateral occipital (LO) complex (MT+/LO complex) (Figure 5b). The right MT

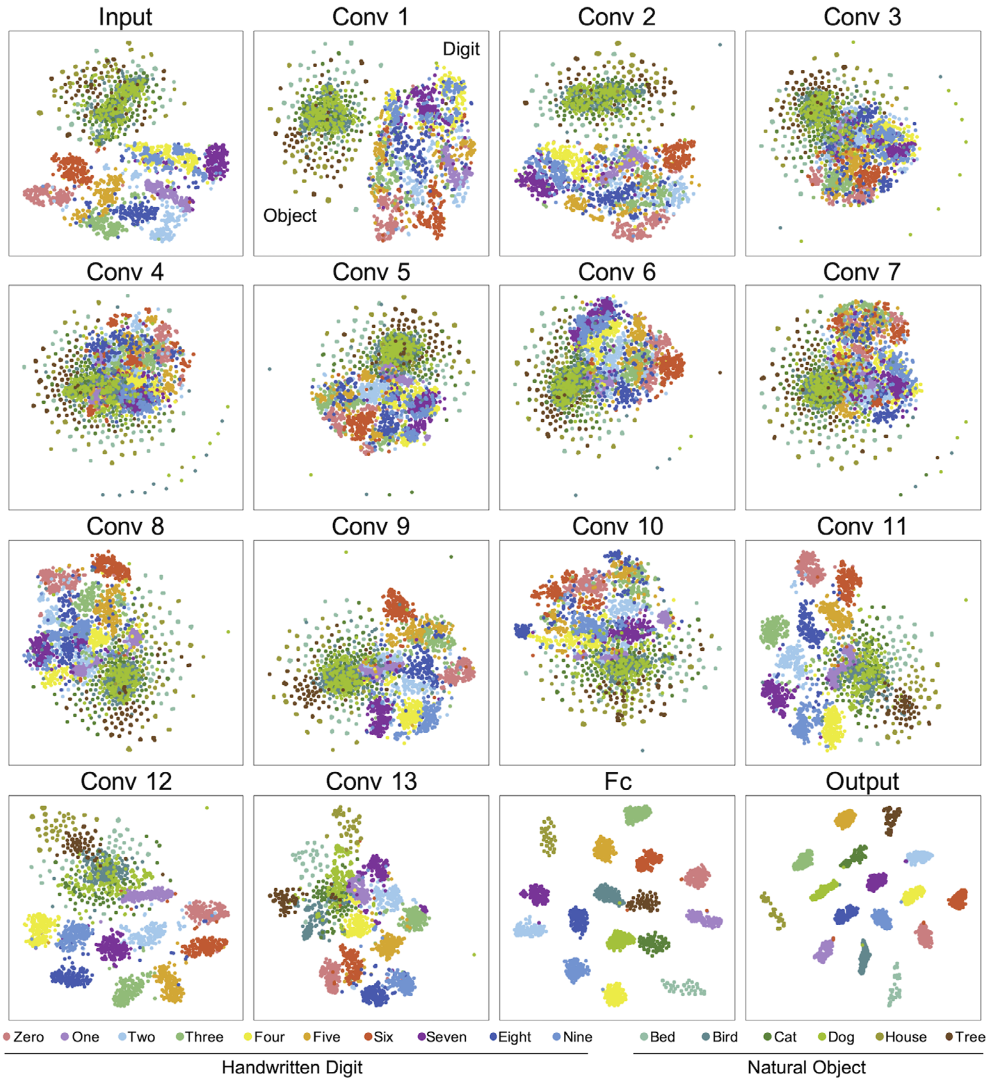

FIGURE 4 *t*-SNE plots of the feature representations for each CNN layer across the 16 classes of handwritten digit and visual object images. Conv, convolutional layer; Fc, fully connected layer; Output, output layer.

+/LO complex was also derived from the two alternative models (i.e., AlexNet and ResNet-50; Figures S15b and S18b). In the intermediate Conv layers, the feature representations had their most significant association mainly with neural activations in the left early visual cortices (EVCs; from the V1 to V4), the bilateral superior parietal cortex (SPC), the posterior cingulate cortex (PCC), and the bilateral dorsal stream visual cortex and bilateral MT+/LO complex (Figure 5c). These areas were also observed for the other two alternative CNN models (Figures S15c and S18c). In the higher Conv layers, the feature representations were strongly associated with higher-level visual association areas such as the MT+/LO complex, ventral stream visual cortex, left inferior parietal cortex, and right SPC (Figure 5d). It is also notable that the right IPC and M1/S1, bilateral medioventral occipital cortex (MVOcC), and left middle cingulate cortex (MCC) were also associated with the higher Conv layers. The association with the relevant brain regions was more distinct in the classification layers (Figure 5e). The association of the MVOcC and MCC with the Fc layer of the CNN was more evident than with Conv 13. The right ventral stream visual cortex, right EVC, left SPC, left orbitofrontal cortex (OFC), and right MPFC exhibited a significant association with the output layer. The degree of similarity across the CNN layers at the group level is summarized for each of the separate clusters in Figures S2–S5. For the two alternative CNN models, higher layers were distinguished from the intermediate layers in the ventral stream visual cortex and IPC in the assignment map (Figures S15d and S18d). The assignment maps of classification layers from the two alternative models were less consistent than the previous layers (Figures S15e and S18e). Nevertheless, the two alternative models and VGG16 included SPC, PCC, and MPFC.

The RSA results using the unsmoothed beta-values presented strikingly similar neural representations compared to the RSA results using the smoothed beta-values (Figures 5 vs. S9). The spatial smoothing of the individual RSA results represented a slightly blurred group inference compared to the individual RSA results without spatial smoothing; however, the overall representations were quite consistent (Figures S9 vs. S11). A hierarchy within the visual areas and across the whole brain, from visual to frontal areas, was commonly observed for the smoothed and unsmoothed individual RSA maps (Figures S9a and S11a). The left EVCs and bilateral stream visual cortex were not only observed in the intermediate Conv layers but also

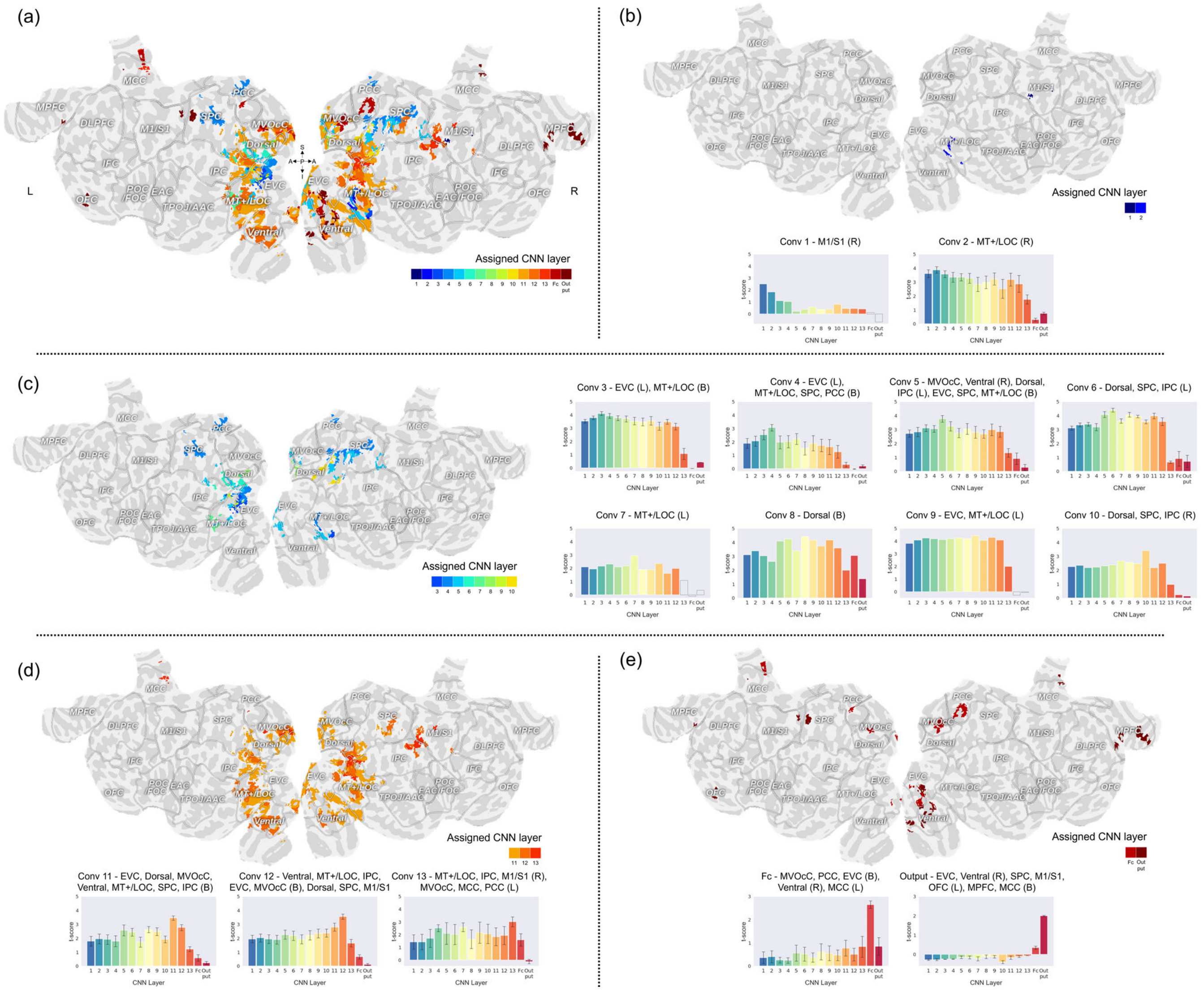

FIGURE 5 CNN layer assignment map and bar graphs indicating the similarity between the neural activations and feature representations for each of the CNN layers. (a) Layer assignment map across all CNN layers. (b) Assignment map for the lower Conv layers (i.e., Conv 1 and 2) and bar graphs of the similarity scores for the ROIs obtained from each of the CNN layers (significant cases, corrected $p < 0.05$ using 5000 random permutations, are color-coded; error bars indicate the standard error across participants). (c) Assignment map and bar graphs of the similarity scores for the intermediate conv layers. (d) Results for the higher conv layers. (e) Results for the classification layers. A, anterior; CNN, convolutional neural networks; Conv, convolutional layer; DLPFC, dorsolateral prefrontal cortex; Dorsal, dorsal stream visual cortex; EAC, early auditory cortex; EVC, early visual cortex; Fc, fully connected layer; I, inferior; IFC, inferior frontal cortex; IPC, inferior parietal cortex; L, left; M1/S1, primary motor cortex and primary somatosensory cortex; MCC, middle cingulate cortex; MPFC, medial prefrontal cortex; MT+/LOC, middle temporal (MT) complex and its neighboring visual areas including lateral occipital (LO) complex; MVOcC, medioventral occipital cortex; OFC, orbitofrontal cortex; Output, output layer; P, posterior; PCC, posterior cingulate cortex; POC/FOC, posterior opercular cortex and frontal opercular cortex; R, right; S, superior; SPC, superior parietal cortex; TPOJ/AAC, temporo-parieto-occipital junction and auditory association cortex; Ventral, ventral stream visual cortex.

in the lower Conv layers when analyzed using unsmoothed beta-values (Figures S9b,c and S11b,c). Most of the areas assigned to the intermediate Conv layers from the smoothed beta-values except the PCC were also observed for the results using unsmoothed beta-values (Figures S9c and S11c). The right EVCs and IPCs were additionally found to be associated with the intermediate Conv layers. In the higher Conv layers, the relevant brain regions, including the higher-order visual association areas, were consistent in addition to the emergence of bilateral DLPFC (Figures S9d and S11d). A bilateral rather than a unilateral association was found in some areas such as MCC and M1/S1. However, brain regions from the classification layers (i.e., Fc and output layers) were rather distinguished from each other depending on the spatial smoothing (Figures S9e and S11e). Although the MVOcC, right ventral stream visual cortex, right EVC, and left OFC showed an association regardless of the smoothing, left MCC and SPC were no longer observed for the unsmoothed beta-values. The right MPFC and bilateral PCC were not found for the smoothed individual RSA maps.

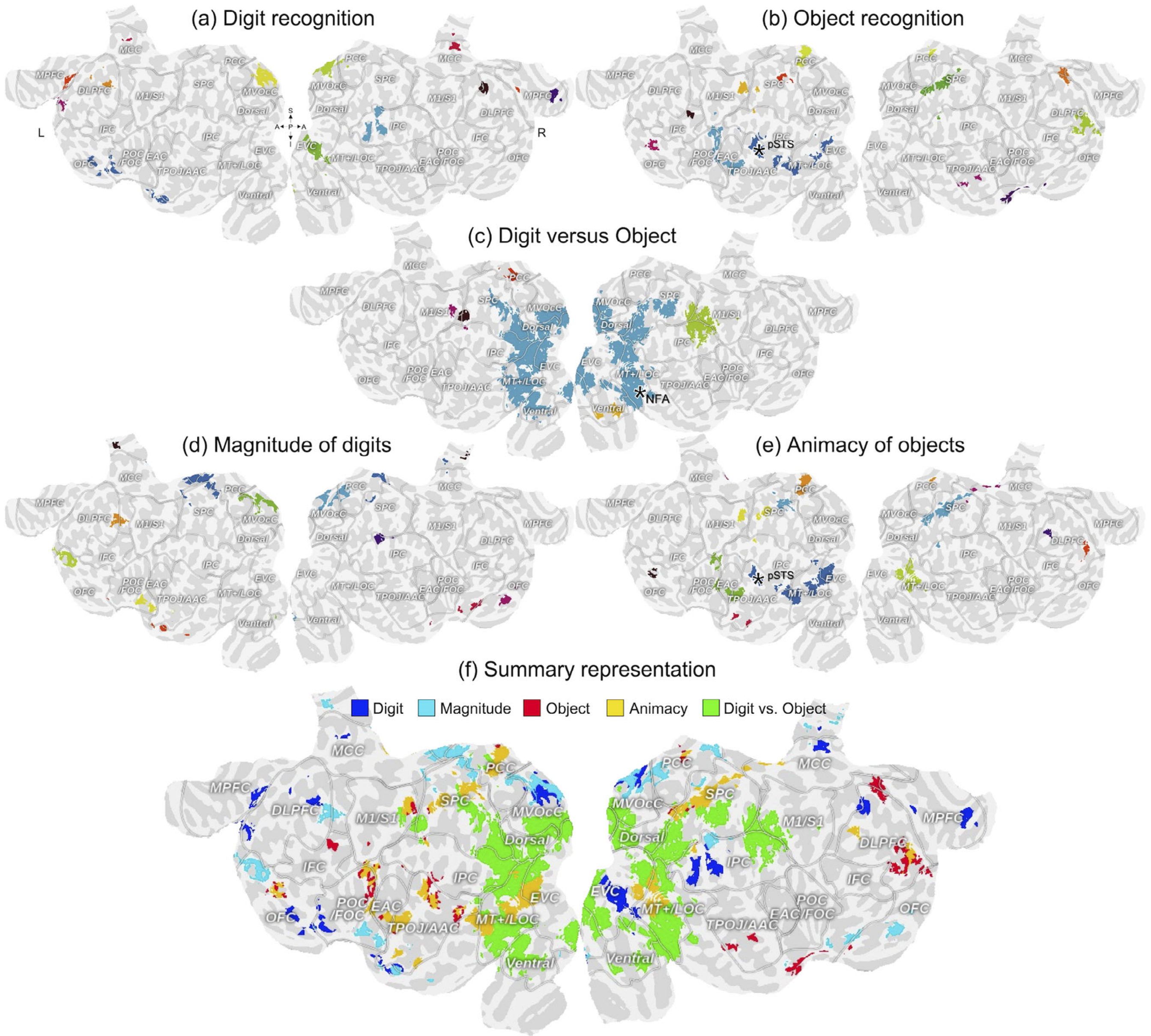

FIGURE 6 Regions-of-interest (ROIs) that encode the visual category/class perception of humans across five different conditions (Figure 1b). In the summary representation, mapping priority is given to the animacy, object, magnitude, digit, and digit versus object conditions in order when there is an overlapping voxel. A, anterior; DLPFC, dorsolateral prefrontal cortex; Dorsal, dorsal stream visual cortex; EAC, early auditory cortex; EVC, early visual cortex; I, inferior; IFC, inferior frontal cortex; IPC, inferior parietal cortex; L, left; M1/S1, primary motor cortex and primary somatosensory cortex; MCC, middle cingulate cortex; MPFC, medial prefrontal cortex; MT+/LOC, middle temporal (MT) complex and its neighboring visual areas including lateral occipital (LO) complex; MVOcC, medioventral occipital cortex; NFA, number form area (Grotheer, Herrmann, et al., 2016); OFC, orbitofrontal cortex; P, posterior; PCC, posterior cingulate cortex; POC/FOC, posterior opercular cortex and frontal opercular cortex; pSTS, posterior superior temporal sulcus; R, right; S, superior; SPC, superior parietal cortex; TPOJ/AAC, temporo-parieto-occipital junction and auditory association cortex; Ventral, ventral stream visual cortex.

3.3 | Neural representations of concepts across categories/classes from humans

Figure 6 presents the brain regions identified from the RSA using the neural activations and the RDM codes that encoded human visual perception. From the handwritten digit recognition, multiple ROIs across the whole brain, such as the visual and associated areas (i.e., right EVC, right IPC, bilateral MVOcC/precuneus), left auditory association cortex (AAC), and frontal areas (i.e., bilateral MPFC/dorsolateral prefrontal cortex [DLPFC]/MCC and left OFC/frontal opercular cortex [FOC]), were identified (Figure 6a). From the visual object recognition, the left EVC, left MT+/LO complex, left EAC, and bilateral PCC were identified, along with the left M1/S1 and bilateral SPC (Figure 6b). In the perception of digits versus objects (Figure 6c), the identified ROIs included the EVC, dorsal/ventral stream visual cortex, MT+/LO complex, IPC, M1/S1, and left PCC, which is markedly similar to the CNN layer

TABLE 1 Detailed information for the regions-of-interest (ROIs; Figure 6a,d) identified for the digit recognition and digit magnitude conditions

Cluster	Size	Foci (*x*, *y*, *z* mm)	*t*-score	Sub-cluster with percentage (%) overlap
Digit recognition				
AAC, FOC, OFC (L)	132	31.5, −25.5, −22.5	5.19	STG (L; 39%), OrG (L; 39%), MTG (L; 5%), IFG (L; 5%), INS (L; 1%)
IPC (R)	103	−37.5, +55.5, +37.5	5.75	IPL (R; 86%), SPL (R; 2%)
EVC (R)	86	−22.5, +88.5, −7.5	8.37	LOcC (R; 81%), FuG (R; 8%), MVOcC (R; 7%)
MVOcC, PCC (R)	82	−28.5, +52.5, +7.5	5.70	MVOcC (R; 51%), Pcun (R; 34%)
MVOcC (L)	68	+25.5, +58.5, +4.5	4.05	MVOcC (L; 69%), Pcun (L; 13%), CG (L; 6%)
DLPFC (L)	46	+31.5, −13.5, +52.5	5.23	MFG (L; 85%)
MPFC (B)	45	+1.5, −43.5, +43.5	5.46	SFG (L; 71%), SFG (R; 20%)
MCC (B)	36	−10.5, +13.5, +49.5	4.30	SFG (R; 28%), PCL (R; 25%), SFG (L; 22%), CG (R; 11%), PCL (L; 6%)
DLPFC, SFG (L)	33	+19.5, −61.5, +25.5	3.62	SFG (L; 58%), MFG (L; 30%)
MPFC (R)	32	−10.5, −37.5, −7.5	4.21	OrG (R; 81%), CG (R; 12%)
DLPFC (R)	25	−28.5, −13.5, +46.5	3.70	MFG (R; 60%), SFG (R; 28%)
Magnitude of digits				
MCC, SPC, PCC (B)	196	−4.5, 37.5, 43.5	2.50	Pcun (L; 51%), CG (R; 19%), Pcun (R; 11%), PCL (L; 8%), SPL (L; 3%), CG (L; 2%), PCL (R; 1%), PoG (L; 1%)
MVOcC, PCC (R)	98	−22.5, +64.5, +4.5	2.34	MVOcC (R; 46%), Pcun (R; 44%)
MVOcC, PCC (L)	72	+25.5, +58.5, +4.5	1.94	Pcun (L; 47%), MVOcC (L; 42%), CG (L; 10%)
OFC (L)	62	+28.5, −52.5, +7.5	2.15	MFG (L; 84%), SFG (L; 5%)
FOC (L)	56	+49.5, −1.5, −10.5	2.17	STG (L; 52%), INS (L; 36%)
DLPFC (L)	51	+34.5, −4.5, +55.5	2.07	MFG (L; 78%), PrG (L; 4%)
AAC, STG (L)	47	+40.5, +1.5, −22.5	2.25	STG (L; 57%), MTG (L; 15%), ITG (L; 4%)
OrG (R)	45	−34.5, −25.5, −13.5	1.87	OrG (R; 76%), STG (R; 7%), INS (R; 2%)
OFC (R)	37	−28.5, −52.5, −13.5	3.04	MFG (R; 46%), OrG (R; 43%)
IPC (R)	31	−37.5, +55.5, +37.5	3.66	IPL (R; 71%), SPL (R; 3%)
CG (B)	20	−7.5, −1.5, +28.5	2.02	CG (L; 50%), CG (R; 25%)

Note: Searchlight-based representational similarity analysis (RSA) was used on the multivoxel patterns of the neural activations and representational dissimilarity matrix (RDM) codes that represent the visual category/class perception of humans. We obtained labels for the (sub-)clusters from the Brainnetome atlas (https://atlas.brainnetome.org).

Abbreviations: AAC, auditory association cortex; B, bilateral; CG, cingulate gyrus; DLPFC, dorsolateral prefrontal cortex; EVC, early visual cortex; FOC, frontal opercular cortex; FuG, fusiform gyrus; IFG, inferior frontal gyrus; INS, insular cortex; IPC, inferior parietal cortex; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; L, left; LOcC, lateral occipital cortex; MCC, middle cingulate cortex; MFG, middle frontal gyrus; MPFC, medial prefrontal cortex; MTG, middle temporal gyrus; MVOcC, medioventral occipital cortex; OFC, orbitofrontal cortex; OrG, orbital gyrus; PCC, posterior cingulate cortex; PCL, paracentral lobule; Pcun, precuneus; PoG, postcentral gyrus; PrG, precentral gyrus; R, right; SFG, superior frontal gyrus; SPC, superior parietal cortex; SPL, superior parietal lobule; STG, superior temporal gyrus.

assignment map (Figure 5a). In the perception of the magnitude of the digits, the bilateral SPC, a middle part of right OFC, and a superior part of left OFC were included when compared to the digit recognition. Bilateral MVOcC, SPC, OFC, and the left temporal pole (in the TPOJ/AAC) were commonly found from both the contrast of [1,2,3] versus [7,8,9] and all the combinatorial groupings of lower and higher digits (i.e., a three-digits subset from [1,2,3,4] vs. a three-digits subset from [6,7,8,9]) (Figure S8). In the perception of the animacy of the objects, it is notable that not only the right MT+/LO complex but also the left superior temporal sulcus (STS) was additionally found when compared to object recognition (Figure 6e). Figure 6f summarizes all the representations, while Tables 1–3 present information for the ROIs.

The representations of human visual perception using unsmoothed beta-values were matched to those of smoothed beta-values (Figures 6, S10, and S12). Among the five RDM codes for human visual perception, the neural representations of digit recognition exhibited the most significant differences in the identified ROIs depending on spatial smoothing of the beta values (Figures S10a and S12a). Visual association areas (including right EVC, right IPC, and bilateral MVOcC), left AAC and the frontal areas (including bilateral MPFC, left DLPFC, and left OFC/FOC) were consistently found to be important. The newly identified areas included visual areas such as left EVC and bilateral ventral/dorsal stream visual cortex, left IPC, and right POC/FOC. The neural representations of object recognition

TABLE 2 Detailed information on the regions-of-interest (ROIs; Figure 6b,e) identified for the object recognition and object animacy conditions

Cluster	Size	Foci (*x*, *y*, *z* mm)	Peak *t*-score	Sub-cluster with percentage (%) overlap
Object recognition				
EVC, MT+/LOC, ITG, TPOJ (L)	262	+55.5, +52.5, +7.5	3.45	LOcC (L; 26%), pSTS (L; 23%), ITG (L; 13%), STG (L; 13%), IPL (L; 13%), MTG (L; 10%)
AAC, EAC, POC (L)	154	+55.5, +19.5, +7.5	3.53	PoG (L; 37%), STG (L; 28%), PrG (L; 18%), INS (L; 16%)
SPC (R)	143	−13.5, +70.5, +52.5	2.38	SPL (R; 48%), LOcC (R; 17%), Pcun (R; 12%), IPL (R; 4%)
DLPFC (R)	118	−25.5, −49.5, +28.5	2.95	MFG (R; 93%)
PCC (B)	83	+1.5, +43.5, +25.5	2.59	CG (L; 51%), Pcun (L; 23%), CG (R; 19%), Pcun (R; 1%)
M1/S1, SPC, IPC (L)	55	+40.5, +37.5, +58.5	1.56	PoG (L; 73%), IPL (L; 5%)
DLPFC, MPFC (R)	49	−19.5, −28.5, +61.5	1.92	SFG (R; 71%)
SPC (L)	40	+4.5, +70.5, +61.5	1.65	SPL (L; 42%), Pcun (L; 12%)
OFC (L)	37	+43.5, −58.5, −7.5	2.40	MFG (L; 78%), OrG (L; 3%)
AAC (R)	35	−58.5, +19.5, −13.5	1.29	MTG (R; 89%), STG (R; 3%)
STG, OFC (R)	29	−28.5, −22.5, −25.5	1.79	STG (R; 66%), INS (R; 10%)
IFC (L)	20	+34.5, −4.5, +25.5	1.79	MFG (L; 55%), PrG (L; 20%), IFG (L; 15%)
Animacy of objects				
EVC, MT+/LOC, ITG, TPOJ (L)	355	+34.5, +97.5, +1.5	3.84	LOcC (L; 49%), ITG (L; 15%), pSTS (L; 11%), MTG (L; 10%), STG (L; 9%), IPL (L; 3%)
SPC (B), IPC (R)	256	−34.5, +79.5, +46.5	2.46	SPL (R; 29%), LOcC (R; 14%), Pcun (R; 13%), SPL (L; 11%), Pcun (L; 9%), IPL (R; 2%)
AAC, EAC, POC (L)	142	+43.5, +19.5, +7.5	2.43	STG (L; 35%), PoG (L; 34%), INS (L; 20%), PrG (L; 11%)
EVC, MT+/LOC (R)	116	−37.5, +91.5, −4.5	3.87	LOcC (R; 99%), MTG (R; 1%)
M1/S1, SPC, IPC (L)	104	+34.5, +52.5, +70.5	1.98	PoG (L; 37%), SPL (L; 20%), IPL (R; 5%)
PCC (B)	77	+4.5, +55.5, +28.5	2.53	CG (L; 56%), Pcun (L; 25%), CG (R; 14%)
DLPFC (R)	32	−25.5, −49.5, +28.5	2.72	MFG (R; 91%)
AAC (L)	30	+55.5, −4.5, −22.5	1.41	MTG (L; 70%), STG (L; 27%)
MCC (L, R)	28	−13.5, +49.5, +64.5	2.03	Pcun (R; 36%), PCL (R; 32%), PCL (L; 14%), SPL (R; 7%)
DLPFC (R)	24	−46.5, −13.5, +55.5	1.63	MFG (R; 58%)
OFC (L)	21	+43.5, −58.5, −7.5	1.68	MFG (L; 81%)

Note: Please refer to Table 1 for more detailed information on the cluster labeling.
Abbreviations: AAC, auditory association cortex; B, bilateral; CG, cingulate gyrus; DLPFC, dorsolateral prefrontal cortex; EAC, early auditory cortex; EVC, early visual cortex; IFC, inferior frontal cortex; IFG, inferior frontal gyrus; INS, insular cortex; IPC, inferior parietal cortex; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; L, left; LOcC, lateral occipital cortex; M1/S1, Primary motor cortex and primary somatosensory cortex; MCC, middle cingulate cortex; MFG, middle frontal gyrus; MPFC, medial prefrontal cortex; MT+/LOC, middle temporal (MT) complex and its neighboring visual area including lateral occipital (LO) complex; MTG, middle temporal gyrus; OFC, orbitofrontal cortex; OrG, orbital gyrus; PCC, posterior cingulate cortex; PCL, paracentral lobule; Pcun, precuneus; POC, posterior opercular cortex; PoG, postcentral gyrus; PrG, precentral gyrus; pSTS, posterior superior temporal sulcus; R, right; SFG, superior frontal gyrus; SPC, superior parietal cortex; SPL, superior parietal lobule; STG, superior temporal gyrus; TPOJ, temporo-parieto-occipital junction.

were similarly observed from ROIs in EVC, MT+/LO complex, M1/S1, SPC, and left EAC (Figures S10b and S12b). The ROIs in the bilateral PCC disappeared, whereas the ones in the bilateral MCC appeared. In the perception of digits versus objects, the associated brain regions were almost identical (Figures S10c and S12c). In the perception of the magnitude of the digits compared to the digit recognition, ROIs in the bilateral SPC were preserved, whereas the ones in the right OFC were not (Figures S10d and S12d). In the perception of the animacy of the objects, the right MT+/LO complex was still found in contrast to object recognition while the left STS disappeared (Figures S10e and S12e).

3.4 | Relationship between the trained CNN model and human visual category/class perception

Figure 7 shows the cosine similarity of the visual category/class perception of humans with each of the CNN layers. Overall, the CNN model had higher similarity with human visual perception as the layers moved toward the output layer except for the categorization of digits versus objects, whose similarity monotonically decreased from lower to higher Conv and classification layers.

Figure 8 summarizes the similarity scores for the neural representations of the ROIs identified from the visual category/class

TABLE 3 Detailed information on the regions-of-interest (ROIs; Figure 6c) identified for the digit versus object condition

Cluster	Size	Foci (*x*, *y*, *z* mm)	Peak *t*-score	Sub-cluster with percentage (%) overlap
Digit versus object				
EVC, MT+/LOC, Ventral, ITG, Dorsal, MVOcC, IPC, SPC (B)	2846	−1.5, +76.5, +4.5	10.52	LOcC (L; 20%), LOcC (R; 17%), MVOcC (R; 11%), MVOcC (L; 10%), FuG (L; 8%), IPL (L; 6%), FuG (R; 5%), IPL (R; 5%), SPL (R; 4%), ITG (L; 2%), ITG (R; 2%), SPL (L; 2%), MTG (L; 1%), MTG (R; 1%)
IPC, SPC, M1/S1 (R)	308	−28.5, +40.5, +52.5	3.95	IPL (R; 41%), PoG (R; 40%), SPL (R; 8%), PrG (R; 1%)
Ventral (R)	79	−22.5, +40.5, −19.5	4.20	FuG (R; 81%), PhG (R; 8%), MVOcC (R; 5%)
PCC (L)	34	+10.5, +52.5, +34.5	3.46	Pcun (L; 65%), CG (L; 32%)
M1/S1, IPC (L)	33	+52.5, +25.5, +58.5	3.35	PoG (L; 70%), IPL (L; 21%)
SPC (L)	30	+40.5, +31.5, +43.5	3.56	IPL (L; 50%), PoG (L; 40%), SPL (L; 7%)

Note: Please refer to Table 1 for more detailed information on the cluster labeling.
Abbreviations: B, bilateral; CG, cingulate gyrus; Dorsal, dorsal stream visual cortex; EVC, early visual cortex; FuG, fusiform gyrus; IPC, inferior parietal cortex; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; ITG, inferior temporal gyrus; L, left; LOcC, lateral occipital cortex; M1/S1, primary motor cortex and primary somatosensory cortex; MT+/LOC, middle temporal (MT) complex and its neighboring visual area including lateral occipital (LO) complex; MTG, middle temporal gyrus; MVOcC, medioventral occipital cortex; PCC, posterior cingulate cortex; Pcun, precuneus; PhG, parahippocampal gyrus; PoG, postcentral gyrus; PrG, precentral gyrus; R, right; SPC, superior parietal cortex; SPL, superior parietal lobule; Ventral, ventral stream visual cortex.

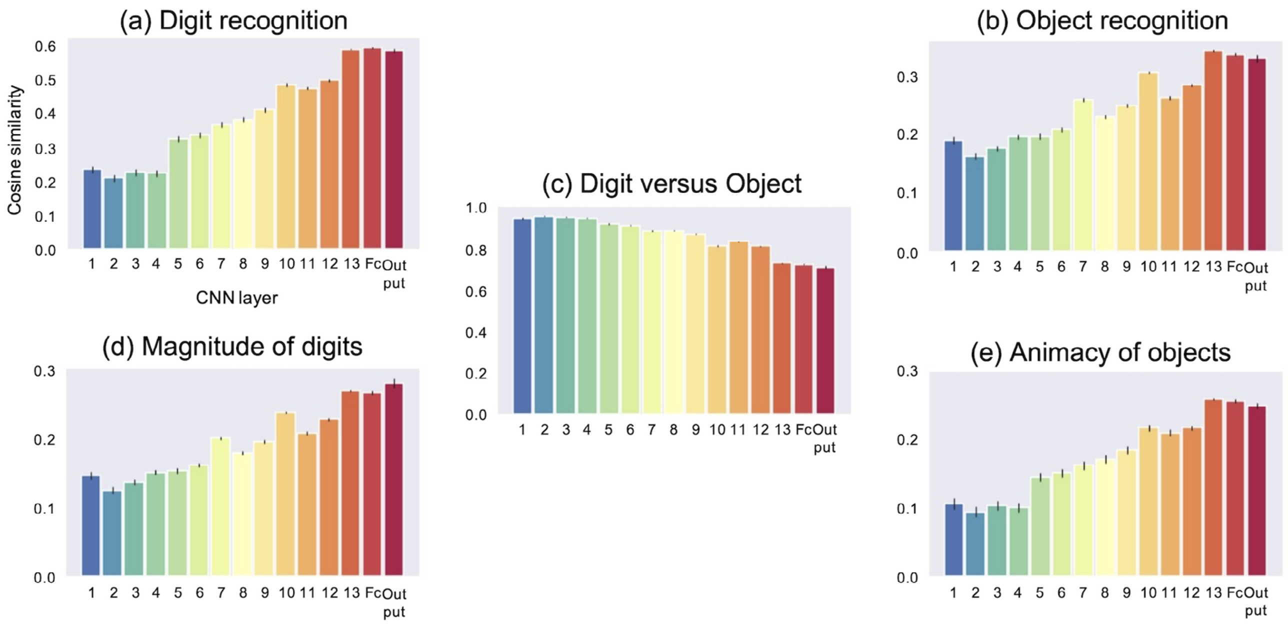

FIGURE 7 Cosine similarity between the RDM for the CNN layers and the RDM for human visual perception across the 16 classes of the digit and object. The mean and standard deviation across participants are illustrated. CNN, convolutional neural networks; Fc, fully connected layer; Output, output layer.

perception of humans and the feature representations for the CNN layers for each of the specific conditions of interest. Despite the broadly similar trends for the specific conditions of interest, the degree of association across the CNN layers for a specific condition varied across the ROIs. Of the ROIs representing digit perception, the right IPC/EVC/DLPFC and bilateral MCC/MPFC exhibited stronger overall associations compared to the AAC and MVOcC (Figure 8a). Generally, a decrease from lower to higher layers was noticeable for the IPC, EVC, MPFC, and DLPFC. Of the ROIs representing object perception, the higher visual areas such as the MT+/LO complex, ITG, and temporo-parieto-occipital junction (TPOJ) demonstrated a strong and consistently positive association with the CNN layers, particularly the intermediate Conv and classification layers (Figure 8b). The PCC, M1/S1, and OFC also exhibited a similar trend. For the digits versus objects condition, the largest cluster, which included the bilateral EVC and associated visual areas across the medial, dorsal, and ventral stream along with the left PCC, showed a strong association with the CNN layers, particularly the intermediate layers (Figure 8c). For the

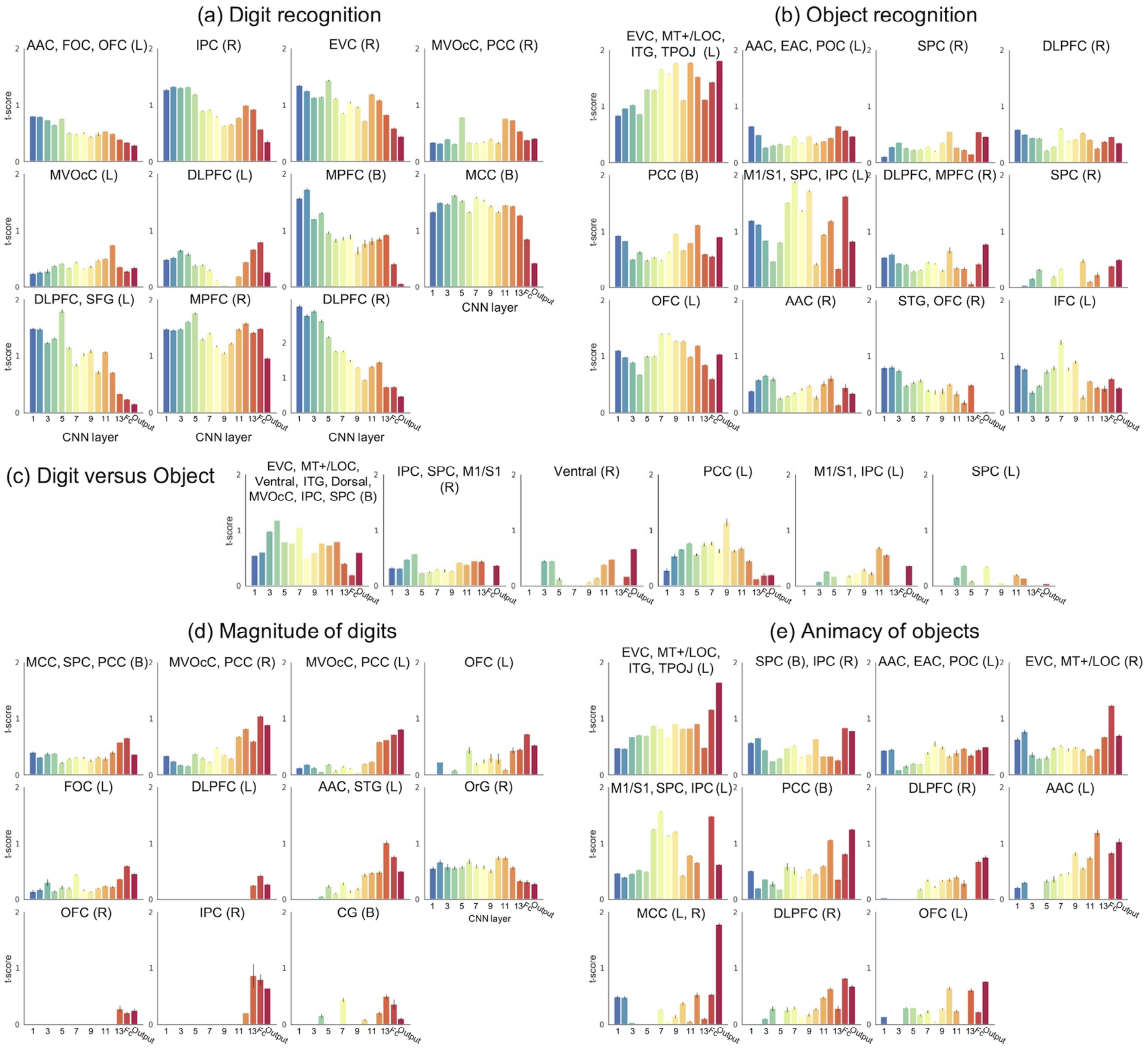

FIGURE 8 RSA using (i) the neural RDM within the ROIs identified from the visual category/class perception of humans and (ii) the RDM for each layer of the trained CNN model. Searchlight RSA is conducted for each of the voxels in the ROI. The voxels with positive *t*-scores from group inference are summarized with their mean *t*-score and standard error of the mean for each ROI. AAC, auditory association cortex; AAC, auditory association cortex; B, bilateral; CG, cingulate gyrus; CNN, convolutional neural networks; DLPFC, dorsolateral prefrontal cortex; Dorsal, dorsal stream visual cortex; EAC, early auditory cortex; EVC, early visual cortex; Fc, fully connected layer; FOC, frontal opercular cortex; IFC, inferior frontal cortex; IPC, inferior parietal cortex; ITG, inferior temporal gyrus; L, left; M1/S1, primary motor cortex and primary somatosensory cortex; MCC, middle cingulate cortex; MPFC, medial prefrontal cortex; MT+/LOC, middle temporal (MT) complex and its neighboring visual areas including lateral occipital (LO) complex; MVOcC, medioventral occipital cortex; OFC, orbitofrontal cortex; OrG, orbital gyrus; Output, output layer; PCC, posterior cingulate cortex; POC, posterior opercular cortex; R, right; SFG, superior frontal gyrus; SPC, superior parietal cortex; STG, superior temporal gyrus; TPOJ, temporo-parieto-occipital junction; Ventral, ventral stream visual cortex.

magnitude of digits, most of the ROIs had a strong connection with the higher Conv and classification layers, including the bilateral MCC/SPC/PCC/MVOcC, right IPC, and bilateral OFC (Figure 8d). For the ROIs of associated with the animacy condition, a strong association with the classification layers of the CNN was particularly evident for the EVC/MT+/LO/TPOJ complex and its neighbors, the M1/S1, SPC, IPC, and MCC (Figure 8e).

4 | DISCUSSION

4.1 | Summary of the study

We investigated the neural representations of handwritten digits and visual objects for a CNN model trained to classify the corresponding visual stimuli. The neural activations for the visual stimuli were measured using fMRI. RSA was used to map the neural representations across the whole brain using feature representations in the CNN layers via holistic geometric space across all the paired classes from digit and/or object categories. Neural representations from the human visual perception of these classes were also obtained using hypothetical RDM codes and compared with the CNN model. We found that the neural representations from the CNN model substantially mapped onto the visual areas and their associated areas, along with a portion of the parietal area (Figure 5a). The representations overlapped mainly with the representations for digit perception in comparison to those for object perception in humans (Figure 6c). In terms of the neural representations from the visual perception of humans, additional regions were mapped across the whole brain, such as the OFC, MPFC, and DLPFC for digit perception and digit magnitude, and the left posterior opercular cortex and frontal opercular cortex (POC/FOC), TPOJ/AAC, and right DLPFC for object perception and animacy (Figures 6 and S6).

4.2 | Interpretation of the neural representations from the CNN model

The VGG16-based CNN model was pre-trained to recognize visual objects and fine-tuned to classify 10 additional handwritten digits. The activation maximization map for each of the 16 output nodes portrayed clear digit and object categories, indicating that the fine-tuned model had successfully learned to recognize novel digit and object stimuli (Figure 3b). As compared to the 99.2% classification accuracy of VGG16, the accuracies of AlexNet and ResNet-50 were slightly lower (i.e., 97.0% and 97.1%, respectively). This might be due to similar hyperparameters (e.g., mini-batch size, learning rate annealing, or dropout rate) of the corresponding models and VGG16. Optimizing hyperparameters of the alternative CNN model may further improve the classification performance. *t*-SNE plots of the feature representations for each of the fine-tuned VGG16-based CNN layers exhibited a hierarchical organization (Figure 4). Interestingly, the initially separated feature representations for the visual objects and handwritten digits (Conv 1 and Conv 2) were intermixed in the intermediate Conv layers. The *t*-SNE representations for digits in the lower Conv layers maintained distinct patterns across all ten digits that were similar to the input features. Numerosity representation in the hand areas has been reported (Anobile et al., 2021) which may explain the maximum similarity between the feature representations in the lower Conv layers and neural representations in the M1/S1. The maximum association between the neural representations in the MT+/LO complex and Conv 2 may be related to differences in the complexity of the image patterns between objects and digits.

The separation between the two sets of categories (i.e., digits vs. objects) appeared in the higher Conv layers (Figure 4). Furthermore, the corresponding feature representations across the 10 digits were more evident, with signs of separation across the six objects, suggesting that the visual features of these 16 classes solidified in the higher Conv layers of the trained CNN. The associated ROIs whose neural representations were maximally associated with the feature representations for the higher Conv layers were mainly identified in the higher-order visual areas such as the dorsal/medial/ventral stream visual association areas and the MT+ and LO complex regions (Figure 5d). The ventral visual stream includes the posterior inferotemporal complex (ITC) and fusiform face complex (FFC) (Glasser et al., 2016). The ITC has been associated with face, object, and scene perception (Conway, 2018) and the ventral part of the ITC has been linked to visual number symbols, such as Arabic numerals (Hannagan et al., 2015). It is also worth noting that there was a significant association between the MT+/LO complex and higher Conv layers. We confirmed that the identified clusters in the MT+/LO complex (Figure 5d) included both the MT+ and LO complex. The MT+ and its neighboring areas, which are known to be sensitive to visual motion, (Gaglianese et al., 2017; Maunsell & Van Essen, 1983) are also associated with the sensitivity to eye movements (Dukelow et al., 2001). This may have impacted this study due to the supposedly high pursuit eye movements for the object condition when compared to the digit condition. The appearance of the LO complex, representing object recognition, (Grill-Spector et al., 2001) is possibly due to the distinct contrast between the objects and digits in the higher Conv layers.

Fully separated feature representations across all digit and object classes were obtained from the classification layers of the CNN (Figure 4). The maximally associated brain regions were mainly identified from the right higher-order visual areas, the ventral stream of the visual association areas, PCC, left SPC, bilateral MCC, and bilateral MPFC (Figure 5e). Notably, the clusters in the PCC, MPFC, and OFC exhibited solid and unique associations with the classification layers when compared with the higher Conv layers (Figure S5), suggesting that these higher-order cognitive areas processed specific conceptual information across the 16 classes. We observed that the left hemisphere showed a prominent hierarchical structure from the lower to the higher layers compared to the right hemisphere. On the right hemisphere, the EVCs were mainly assigned to the classification layers (Figure 5e). One potential cause may be due to the difference of the low-level features between the digits and objects (i.e., line-drawings for the digits and color images with background scene for objects). Thus, the EVCs may be highly associated with the feature representation of higher CNN layers. Alternatively, it might be due to a large individual variability. When we inspected the association with CNN layers for each of the 15 subjects rather than the group result, many subjects showed an assignment in the EVCs, with the lower to intermediate layers (Figure S7). However, the varying spatial locations in the EVC across the subjects might have failed to reflect in the group inference. Interestingly, the association of the EVC with the classification layers was varying depending on whether the spatial smoothing was applied before the *t*-test or not (Figures S9 and S11), and was not evident from an alternative CNN model, ResNet-50.

When the results were obtained using unsmoothed beta-valued maps, the layer assignment map on the whole brain was remarkably similar, mostly preserved in the higher layers but the difference was mainly found in the early layers (Figures 5 vs. S9 and S11). We attribute the change of the whole-brain map from the searchlight RSA to spatially noisy beta-valued maps when the spatial smoothing was not applied. The searchlight area included voxels in the range of three-voxel radius sphere (i.e., the maximum distance of 12 mm within a sphere) to obtain multivoxel pattern of neural activations. Thus, an application of spatial smoothing using Gaussian kernel with an 8 mm FWHM might have substantially altered the neural RDM.

There was an overall similarity of the RSA mapping using two sets of VGG16 models for 16 classes classification that was trained using two sets of six object classes. Specifically, there was the relatively large mapping of brain regions from the higher Conv layer and classification layers in comparison to the lower and intermediate Conv layers (Figures 5 vs. S21).

4.3 | Interpretation of the neural representations from human visual category/class perception

4.3.1 | Digits and their magnitude

ROIs were found bilaterally in the MVOcC overlapping with the anterior part of V1, not only from the digit recognition but also from their magnitude. Notably, the ROIs were retained from the magnitude contrast when alternative groups of lower and higher numbers were used to define a magnitude. In this analysis, the complexity of visual features may also be different between the lower ([1, 2, 3]) and higher ([7, 8, 9]) magnitude groups. However, the complexity of visual feature of the digit “4” is relatively more complex than the visual features of “1” or “2” and seems comparable to that of digits “6,” “7,” or “9.” Thus, our analysis using the groupings of the lower ([1, 2, 3, 4]) and higher ([6, 7, 8, 9]) digits might have alleviated this potential confounding issue on visual feature complexity for the magnitude comparison condition. Nonetheless, the finding of the primary visual cortex in the magnitude perception, across our adopted groupings of numbers, may suggest that there is a potential difference in visual feature complexity between the groups of the high-magnitude and low-magnitude numbers.

We also identified the DLPFC, which is potentially related to number processing (Nieder, 2016). In our experimental setting, participants had to consistently remain attentive to the digits in order to press the button when the presented digit was “0.” Thus, the cognitive process required to recognize “0” from among the set of ten digits may produce neural activations in the frontal areas, including the MPFC and DLPFC, in the digit recognition condition (Figure 6a).

The regions associated exclusively with the magnitude perception, and not the digit recognition, included the SPC and a superior part of the left OFC across various sets of groupings of digit magnitudes. The intraparietal sulcus (IPS), the horizontal area dividing the IPC and SPC, has been reported to be a numeric information-processing region in both human and nonhuman primates (Ashkenazi et al., 2008; Eger et al., 2003; Isaacs et al., 2001; Nieder, 2016; Piazza et al., 2007; Vallentin et al., 2012). Similarly, the perception of numerical magnitude or quantity is mainly associated with the right IPS, particularly when Arabic numerals are used (Ansari et al., 2007; Arsalidou & Taylor, 2011). The IPS is the first region to process number information in the neural number network that includes the parietal region and the lateral prefrontal cortex (Nieder, 2016). Given the results of previous studies, human perception of both digit recognition and the magnitude of digits explain the association with the IPS in our findings. In another study, the inferior parietal lobule (BA 40) was significantly associated with number calculation tasks in children, extending further to parts of the inferior parietal sulcus in the right hemisphere and the parietal cortex in the left hemisphere (Arsalidou et al., 2018). The ROI in the superior part of the left OFC (as we defined the boundary) was adjacent to the prefrontal cortex, which was reported to have association neurons for numerical symbols in rhesus monkeys (Diester & Nieder, 2007).

4.3.2 | Objects and their animacy

Although the object perception and animacy condition codes had very similar representations overall, the right MT+ and LO complex were identified for the object animacy condition but not with the object recognition condition (Figure 6b,e). This may be because of the cluster in the MT+ complex, which has been associated with visual motion (Gaglianese et al., 2017; Maunsell & Van Essen, 1983). The identification of the MT+ area only in the animacy condition may be because of visual motion imagery (Goebel et al., 1998) because the participants may have evoked motion imagery during the perception of the animate objects (i.e., bird, cat, and dog) but not with the inanimate objects (i.e., bed, house, and tree).

It is widely known that the V5/MT+ area is associated with a general motion, whereas the posterior part of STS is more selective to a biological motion (Beauchamp, 2015; Beauchamp et al., 2003; Deen et al., 2015; Pelphrey et al., 2003). The images of the classes for animacy condition (i.e., bird, cat, and dog) may derive biological motions, which justifies the observation in STS for the animacy perception of objects.

4.4 | Digits versus objects

The ROIs identified from the RDM code for visual category perception contrasting digits and objects (Figure 6c) were strikingly similar to the CNN layer assignment map (Figure 5a), particularly with the brain regions assigned from the Conv layers (Figure 5b–d). Most of the brain regions assigned from the classification layers of the CNN (Figure 5e) were not identified in the ROIs relevant to the visual perception of digits versus objects (Figure S6a,b). This finding was consistent with the *t*-SNE plots of the feature representations for the Conv layers. Each of the 16 classes across the digits and objects showed distinct representations only in the classification layers (Figure 4).

The sensorimotor numerosity system is located close to neural representations associated with hand/digit movement related to counting numbers and numerosity representation, which enables the mediation of the psychophysical interaction between the two systems

(Anobile et al., 2021). We also found clusters in the M1/S1 areas, including parts of the SPC/IPC, possibly due to the contrast between digit and object recognition (Figure 6c). Many studies have demonstrated that the visual number form area is located in the right pITG, (Daitch et al., 2016; Grotheer et al., 2018; Grotheer, Ambrus, & Kovács, 2016; Grotheer, Herrmann, & Kovács, 2016; Shum et al., 2013; Yeo et al., 2017, 2020) which was confirmed by our neural representations (Figure 6c).

We used gray-scale images for the digit category and color images for the object category. The primary and higher-order visual areas observed in our findings (Figure 6c) may be related to the color perception of the human brain, which includes a color-processing stream that begins from the retina to the early visual areas such as the V1 and V2 (Shapley & Hawken, 2011; Zeki & Marini, 1998) and the higher-order visual areas including the V4 and V8 areas (Grill-Spector & Malach, 2004; Hadjikhani et al., 1998).

4.4.1 | Unsmoothed beta-values

The representation of human visual perception using unsmoothed beta-values were noisier than the representation using smoothed beta-values; however, they presented substantially similar locations of ROIs (Figures S10 and S12). Particularly, the representations from the visual category perception contrasting digits to objects were mostly consistent.

4.4.2 | Generalization on a new set of six objects from another dataset

When neural RDMs were constructed using beta-valued brain maps of bear, beetle, car, flower, monkey, and pen in BOLD5000, the neural representations were partly reproduced but showed a large difference (Figures S22 vs. 6). Overall, a substantial association was found in the MPFC and the anterior to middle part of STS. The TPOJ and OFC were commonly found for the object recognition condition (Figures 6b and S22b). The right DLPFC was observed in common for the perception of animacy (Figures 6e and S22c). For the visual perception of digits versus objects, the ROIs spanned the whole brain including the replicated findings in EVC, MT+/LO complex, dorsal/ventral stream visual cortex, MVOcC, PCC, SPC, and M1/S1 (Figures 6c and S22a). It was notable that the discrepancy of the RSA mapping using the RDM codes for human perception of the object recognition and animacy although the RSA mapping of digits versus objects was commonly included the visual and its associative areas. This could be due to the limitation of our alternative RSA which used the BOLD5000 dataset for the object classes while our dataset was used for the digit classes. Our core assumption was that the neural RDM constructed with heterogenous subjects would maintain the holistic geometric representations across the 16 classes. However, the various confounding factors such as heterogeneities of MRI scanners, imaging parameters, and preprocessing options might have caused the less degree of replication between the two sets of the RSA mapping results. Future study is warranted in this context.

4.5 | Interpretation of the ROIs for human visual perception and from the CNN model

The cosine similarity between the RDM for human category/class perception and the RDM for the CNN layers monotonically increased as the layers became deeper (Figure 7). This suggests that the CNN model classified the information on the digits and their magnitude and on the objects and their animacy hierarchically from the lower to higher Conv layers. The categorization between digits and objects seemed to be established relatively early in the lower Conv layers. Fine-grained information was obtained from the similarity between the neural RDM for the ROIs from the category/class perception of humans and the RDM for the CNN layers (Figure 8). The generally decreasing pattern of the similarity from the ROIs related to digit recognition (which is similar to the trend shown in Figure 7c) might be possible because these ROIs were identified from the main contrast of digits (dissimilar across digits) versus objects (similar across objects) (bottom of Figure 1b). The ROIs for object recognition in humans appeared to be more strongly connected to the object recognition in the CNN model, as evidenced by the strong association for the EVC/MT+/LO complex and its neighboring areas (Figure 8b) (Grill-Spector et al., 2001).

The ROIs identified from the distinction between digits and objects had a high similarity with the feature representations in the intermediate and higher Conv layers, particularly for the EVC and dorsal/medial/ventral visual areas, PCC, M1/S1, and IPC (Figure 8c). The neural representations for human visual category/class perception (Figure S6c) were more widely distributed across the whole brain than was the CNN layer assignment map (Figure S6a). Notably, the DLPFC, OFC, IFC, POC/FOC, and TPOJ/AAC had a significant association with human visual category/class perception. At the same time, many of these representations were obscured in the CNN, possibly because this model only processes visual stimuli-related information from the input images to maximize classification performance in the output layer. On the other hand, the RDM codes that imitated human perception may have enabled the identification of neural representations for explicit information of a class and implicit high-level cognitive processes relevant to the condition of interest.

4.6 | Utility of brain-encoding studies using neural network models

Neural network-based models are capable of state-of-the-art performance in object recognition, with a CNN outperforming alternative models for 1000 visual object recognition tasks in 2012 (Krizhevsky et al., 2012) and surpassing human-level performance in a visual object recognition task (He et al., 2015). It appears that no other computational models are capable of outperforming deep neural network

(DNN)-based models in sensory perception tasks, which suggests that DNN models are valuable computational models for understanding human brain perception (Kell & McDermott, 2019). A growing body of research has utilized DNN models to understand neural representations within the human brain as measured using non-invasive fMRI (Cross et al., 2020; Jain & Huth, 2018; Kell et al., 2018). Investigations of brain-encoding models have also extended their focus from visual perception to sound/auditory perception (Kell et al., 2018), sentiment analysis using language processing models (Jain & Huth, 2018), and higher-order cognitive tasks (Saxe et al., 2020). Neural representations of the action and decision-making processes of humans during game-playing scenarios have also been investigated using a deep reinforcement learning model (Cross et al., 2020). However, artificial general intelligence using DNN-based computational models is still in its infancy, and there is significant room for improvement in imitating human-level intelligence (Jordan, 2019). We believe that brain-encoding research identifying neural representations from computational models for cognitive processes, including sensory information perception and the systematic comparison with human perception, provides an invaluable systematic framework for the development of human-inspired computational models (Hassabis et al., 2017). For example, similarly to our analyses to compare the CNN models (i.e., AlexNet, ResNet-50, and VGG16) with the human visual perception, we can also conduct the similarity mapping between computational models and cognitive processes of the human brain. Consequently, network architectures and/or hyperparameters can be adjusted to find the most suitable computational models that resemble the information processing of the human brain.

4.7 | Potential weaknesses and future work

Our hypothetical RDM codes for the visual category/class perception of humans may inherently be limited in terms of accommodating a broad range of human perception across our adopted handwritten digit and visual object classes. Constructing the codes for human perception based on the participants' behavioral data may provide a rich set of perceptual codes that account for higher-level cognitive processes (Bracci et al., 2019; Contini et al., 2020; Kim et al., 2020; King et al., 2019). Our CNN model was trained to recognize all 16 digit and object classes in a single output layer. Alternatively, dual-task CNN architecture used to distinguish speech perception from music perception (Kell et al., 2018) can be gainfully employed in our experimental setting for the visual perception of digits and objects. It would be interesting for future work to systematically compare the neural representations for digit and object perception across various architectures for CNN-based computer vision models. An extended analysis using alternative CNN-based classification models was conducted using AlexNet and ResNet-50, in addition to VGG16. We noticed that even if the architecture and the number of layers were different across the CNN models, the hierarchical neural representations across the layers were markedly similar (Figures 5 and S15 and S18). This may suggest that our findings can be generalizable to CNN-based models which is warranted in a future study.

We used handwritten digits and visual objects as visual stimuli. Future studies can thus investigate the neural representations for visual perception for (a) handwritten digits compared to typed digits and (b) natural objects compared to artificial objects by collecting additional datasets. The low-level visual difference between digits and objects (e.g., drawing vs. picture; background) could limit the interpretation of our findings, which showed a considerable association in the early visual processing areas. Thus, it would be worthwhile to perform the similar analyses using closely matched visual stimuli using line-drawings such as the Google's QuickDraw (https://quickdraw.withgoogle.com/; https://githubd.com/googlecreativelab/quickdraw-dataset) or sketch of the visual scene (Lee et al., 2021, 2022).

Symbolic characters/letters are the building block of language. The associated neural representations can be investigated using CNN-based computational models to recognize characters/alphabets and associated neural activations. The visual pattern recognition of characters/letters can be further extended to understanding the neural representations of words based on visual perception. Alternatively, the words can be represented as embedding vectors based on the contextual information of the words included in a set of sentences (Devlin et al., 2018; Mikolov et al., 2013; Pennington et al., 2014). Embedding vectors of words have been gainfully employed to understand the contextual information of sentences using a DNN with LSTM units (Jain & Huth, 2018). It is important for future research to systematically compare the neural representations from (a) neural net-based models that can recognize the meaning of words based on visual stimuli-driven perception and (b) neural net-based models that can recognize the meaning of words based on their numerically embedded vectors. This investigation may provide insight into the construction of an optimal computational model to understand words that stem from the associated human perception. We can also extend this line of study to the neural representations of sentences, which may benefit the development of computational models that can better understand the concrete and contextual meaning of sentences.

The dissimilarity metric to define the neural RDM and the RDM of the CNN layers (or, human visual perception) was calculated based on 1—Pearson's correlation coefficient. Alternatively, Euclidean distance or Mahalanobis distance can be adopted as a dissimilarity measure (Kriegeskorte et al., 2008), which may alter the RSA results. For example, the absolute difference of activations largely affects the dissimilarity value when the Euclidean distance is used. The reliability can also be affected by the dissimilarity metrics (Walther et al., 2016). We set the radius of a spherical searchlight area as three-voxel size. Alternatively, a two-voxel size radius is also often adopted for the searchlight multivoxel patterns. For example, Bulthé and colleagues used two-voxel radius for multi-voxel pattern analysis to represent the symbolic numbers in the cortex (Bulthé et al., 2014). The different size of radius for searchlight analysis was also compared, in which a smaller searchlight radius provides a greater spatial selectivity than a larger radius (Coutanche et al., 2011; Oosterhof et al., 2011). We also conducted the RSA using two-voxel size radius instead of three, however, the RSA result was comparable to the RSA result of three-voxel radius (data not shown). The explained variance criterion for dimensionality reduction (i.e., 90% in our study) is another hyperparameter that may alter the results. The potential variability of the RSA results depending on these hyperparameters warrants a future investigation.

5 | CONCLUSION

We investigated the neural representations of visual perception across ten handwritten digits and six visual objects from two distinct perspectives: (a) a CNN model that almost perfectly recognized the 16 classes and (b) the hypothetical visual category/class perception of humans. Digit perception is an essential process for the human brain in terms of visual symbolic letter processing. Object perception is also crucial for understanding straightforward concepts in a visual scene. Our findings suggest that our adopted CNN model successfully reflected visual category/class perception across heterogeneous stimuli (i.e., abstract symbols and concrete objects). However, the corresponding neural representation was lacking, particularly in the higher cognitive processing areas of the human brain across the prefrontal, parietal, and temporal regions. The example dataset and the codes used to analyze the dataset and visualize the results are publicly available from our GitHub repository (https://github.com/bsplku/ISL_RSA_DigitObject_Visual). Recognition of visual stimuli, including symbolic digits and concrete objects, is crucial to understanding the core concepts of a visual scene, particularly for use in, for example, image caption models. We could utilize an ecologically motivated image dataset (Mehrer et al., 2020) to better understand an image caption model in comparison with the human brain. We believe that investigating the neural representations from computer vision models and their applications, such as in image captioning systems, would provide valuable insight into the development of human-inspired computational models.

ACKNOWLEDGMENTS

The authors thank Changha Lee and Wangwon Lee for their logistic support during data collection, and Minseok Choi for helpful discussion and visualization support during the manuscript preparation.

FUNDING INFORMATION

This work was supported by the National Research Foundation (NRF) grant, Ministry of Science and ICT (MSIT) of Korea (NRF-2017R1E1A1A01077288, 2021M3E5D2A01022515), and in part by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (22ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System).

CONFLICT OF INTEREST

The authors have no conflicts of interest regarding this study, including financial, consultant, institutional, or other relationships. The sponsor was not involved in the study design, data collection, analysis or interpretation of the data, manuscript preparation, or the decision to submit for publication.

DATA AVAILABILITY STATEMENT

The example dataset and the codes used to analyze the dataset and visualize the results are publicly available from our GitHub repository (https://github.com/bsplku/ISL_RSA_DigitObject_Visual).

ORCID

Jong-Hwan Lee https://orcid.org/0000-0002-8902-6009

REFERENCES

Anobile, G., Arrighi, R., Castaldi, E., & Burr, D. C. (2021). A sensorimotor numerosity system. *Trends in Cognitive Sciences*, *25*(1), 24–36. https://doi.org/10.1016/j.tics.2020.10.009

Ansari, D., Lyons, I. M., van Eimeren, L., & Xu, F. (2007). Linking visual attention and number processing in the brain: The role of the temporo-parietal junction in small and large symbolic and nonsymbolic number comparison. *Journal of Cognitive Neuroscience*, *19*(11), 1845–1853.

Arsalidou, M., Pawliw-Levac, M., Sadeghi, M., & Pascual-Leone, J. (2018). Brain areas associated with numbers and calculations in children: Meta-analyses of fMRI studies. *Developmental Cognitive Neuroscience*, *30*, 239–250. https://doi.org/10.1016/j.dcn.2017.08.002

Arsalidou, M., & Taylor, M. J. (2011). Is 2 + 2 = 4? Meta-analyses of brain areas needed for numbers and calculations. *NeuroImage*, *54*(3), 2382–2393. https://doi.org/10.1016/j.neuroimage.2010.10.009

Ashkenazi, S., Henik, A., Ifergane, G., & Shelef, I. (2008). Basic numerical processing in left intraparietal sulcus (IPS) acalculia. *Cortex*, *44*(4), 439–448. https://doi.org/10.1016/j.cortex.2007.08.008

Beauchamp, M. S. (2015). The social mysteries of the superior temporal sulcus. *Trends in Cognitive Sciences*, *19*(9), 489–490.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. *Journal of Cognitive Neuroscience*, *15*(7), 991–1001.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101.

Bracci, S., Ritchie, J. B., Kalfas, I., & Op de Beeck, H. P. (2019). The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience*, *39*(33), 6513–6525.

Bulthé, J., De Smedt, B., & Op de Beeck, H. P. (2014). Format-dependent representations of symbolic and non-symbolic numbers in the human cortex as revealed by multi-voxel pattern analyses. *NeuroImage*, *87*, 311–322. https://doi.org/10.1016/j.neuroimage.2013.10.049

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 1–13. https://doi.org/10.1038/srep27755

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature. Communications*, *11*(1), 1–13. https://doi.org/10.1038/s41467-020-14578-5

Contini, E. W., Goddard, E., Grootswagers, T., Williams, M., & Carlson, T. (2020). A humanness dimension to visual object coding in the brain. *NeuroImage*, *221*, 117139. https://doi.org/10.1016/j.neuroimage.2020.117139

Conway, B. R. (2018). The organization and operation of inferior temporal cortex. *Annual Review of Vision Science*, *4*(1), 381–402. https://doi.org/10.1146/annurev-vision-091517-034202

Coutanche, M. N., Thompson-Schill, S. L., & Schultz, R. T. (2011). Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage*, *57*(1), 113–123.

Cross, L., Cockburn, J., Yue, Y., & O'Doherty, J. P. (2020). Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron*, *109*, 724–738.e7. https://doi.org/10.1016/j.neuron.2020.11.021

Daitch, A. L., Foster, B. L., Schrouff, J., Rangarajan, V., Kaşikçi, I., Gattas, S., & Parvizi, J. (2016). Mapping human temporal and parietal neuronal population activity and functional coupling during mathematical cognition. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(46), E7277–E7286. https://doi.org/10.1073/pnas.1608434113

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, *25*(11), 4596–4609.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1), 1–42. https://doi.org/10.1016/0010-0277(92)90049-N

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition*, *1*(1), 83–120.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, *2009*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint*, ArXiv 1810.04805.

Diester, I., & Nieder, A. (2007). Semantic associations between signs and numerical categories in the prefrontal cortex. *PLoS Biology*, *5*(11), e294.

Dukelow, S. P., DeSouza, J. F., Culham, J. C., van den Berg, A. V., Menon, R. S., & Vilis, T. (2001). Distinguishing subregions of the human MT+ complex using visual fields and pursuit eye movements. *Journal of Neurophysiology*, *86*(4), 1991–2000.

Eger, E., Sterzer, P., Russ, M. O., Giraud, A.-L., & Kleinschmidt, A. (2003). A supramodal number representation in human intraparietal cortex. *Neuron*, *37*(4), 719–725. https://doi.org/10.1016/s0896-6273(03)00036-9

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. Technical Report, Univeristé de Montréal.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The human Brainnetome atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, *26*(8), 3508–3526. https://doi.org/10.1093/cercor/bhw157

Gaglianese, A., Vansteensel, M. J., Harvey, B. M., Dumoulin, S. O., Petridou, N., & Ramsey, N. F. (2017). Correspondence between fMRI and electrophysiology during visual motion processing in human MT+. *NeuroImage*, *155*, 480–489.

Gao, J. S., Huth, A. G., Lescroart, M. D., & Gallant, J. L. (2015). Pycortex: An interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, *9*(23), 84. https://doi.org/10.3389/fninf.2015.00023

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), Article 7615–Article 7178. https://doi.org/10.1038/nature18933

Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., & Singer, W. (1998). The constructive nature of vision: Direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. *European Journal of Neuroscience*, *10*(5), 1563–1573.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10), 1409–1422. https://doi.org/10.1016/S0042-6989(01)00073-6

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, *27*, 649–677.

Grotheer, M., Ambrus, G. G., & Kovács, G. (2016). Causal evidence of the involvement of the number form area in the visual detection of numbers and letters. *NeuroImage*, *132*, 314–319. https://doi.org/10.1016/j.neuroimage.2016.02.069

Grotheer, M., Herrmann, K.-H., & Kovács, G. (2016). Neuroimaging evidence of a bilateral representation for visually presented numbers. *Journal of Neuroscience*, *36*(1), 88–97.

Grotheer, M., Jeska, B., & Grill-Spector, K. (2018). A preference for mathematical processing outweighs the selectivity for Arabic numbers in the inferior temporal gyrus. *NeuroImage*, *175*, 188–200. https://doi.org/10.1016/j.neuroimage.2018.03.064

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hadjikhani, N., Liu, A. K., Dale, A. M., Cavanagh, P., & Tootell, R. B. (1998). Retinotopy and color sensitivity in human visual cortical area V8. *Nature Neuroscience*, *1*(3), 235–241.

Hannagan, T., Amedi, A., Cohen, L., Dehaene-Lambertz, G., & Dehaene, S. (2015). Origins of the specialization for letters and numbers in ventral occipitotemporal cortex. *Trends in Cognitive Sciences*, *19*(7), 374–382. https://doi.org/10.1016/j.tics.2015.05.006

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Isaacs, E. B., Edmonds, C. J., Lucas, A., & Gadian, D. G. (2001). Calculation difficulties in children of very low birthweight: A neural correlate. *Brain*, *124*(9), 1701–1707. https://doi.org/10.1093/brain/124.9.1701

Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fmri. *Advances in Neural Information Processing Systems*, *31*, 6628–6637.

Jordan, M. I. (2019). Artificial intelligence—The revolution hasn't happened yet. *Harvard Data Science Review*, *1*(1), 1–9.

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.

Kim, H.-C., Jin, S., Jo, S., & Lee, J.-H. (2020). A naturalistic viewing paradigm using 360° panoramic video clips and real-time field-of-view changes with eye-gaze tracking. *NeuroImage*, *216*, 116617. https://doi.org/10.1016/j.neuroimage.2020.116617

King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J., & Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, *197*, 368–382. https://doi.org/10.1016/j.neuroimage.2019.04.079

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. *ArXiv Preprint*, ArXiv:1412.6980 [Cs]. http://arxiv.org/abs/1412.6980

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Lee, J., Kim, H.-C., Kim, J., Jo, S., Jung, M., & Lee, J.-H. (2021). Brain information processing in different modalities: From activation to connectivity. *Organization of Human Brain Mapping*.

Lee, J., Kim, H.-C., & Lee, J.-H. (2022). Multimodal image and text processing of human brain using artificial neural networks and fMRI. *Organization of Human Brain Mapping*.

Lindh, D., Sligte, I. G., Assecondi, S., Shapiro, K. L., & Charest, I. (2019). Conscious perception of natural images is constrained by category-related visual features. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-12135-3

Maunsell, J. H., & Van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *Journal of Neurophysiology*, *49*(5), 1127–1147.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118. https://doi.org/10.1073/pnas.2011417118

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature. Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-19632-w

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Nieder, A. (2016). The neuronal code for number. *Nature Reviews Neuroscience*, *17*(6), 366–382. https://doi.org/10.1038/nrn.2016.40

Nieder, A. (2021). The evolutionary history of brains for numbers. *Trends in Cognitive Sciences*, *25*(7), 608–621. https://doi.org/10.1016/j.tics.2021.03.012

Oosterhof, N. N., Wiestler, T., Downing, P. E., & Diedrichsen, J. (2011). A comparison of volume-based and surface-based multi-voxel pattern analysis. *NeuroImage*, *56*(2), 593–600.

Pelphrey, K. A., Mitchell, T. V., McKeown, M. J., Goldstein, J., Allison, T., & McCarthy, G. (2003). Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. *Journal of Neuroscience*, *23*(17), 6819–6825.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Piazza, M., Pinel, P., Bihan, D. L., & Dehaene, S. (2007). A magnitude code common to Numerosities and number symbols in human intraparietal cortex. *Neuron*, *53*(2), 293–305. https://doi.org/10.1016/j.neuron.2006.11.022

Saxe, A., Nelli, S., & Summerfield, C. (2020). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67. https://doi.org/10.1038/s41583-020-00395-8

Shapley, R., & Hawken, M. J. (2011). Color in the cortex: Single- and double-opponent cells. *Vision Research*, *51*(7), 701–717. https://doi.org/10.1016/j.visres.2011.02.012

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), e1006633.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Shum, J., Hermes, D., Foster, B. L., Dastjerdi, M., Rangarajan, V., Winawer, J., Miller, K. J., & Parvizi, J. (2013). A brain area for visual numerals. *Journal of Neuroscience*, *33*(16), 6709–6715. https://doi.org/10.1523/JNEUROSCI.4558-12.2013

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint*, ArXiv:1409.1556 [Cs]. http://arxiv.org/abs/1409.1556

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, *106*(31), 13040–13045.

Vallentin, D., Bongard, S., & Nieder, A. (2012). Numerical rule coding in the prefrontal, premotor, and posterior parietal cortices of macaques. *Journal of Neuroscience*, *32*(19), 6621–6630. https://doi.org/10.1523/JNEUROSCI.5071-11.2012

Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136–4160. https://doi.org/10.1093/cercor/bhx268

Yeo, D. J., Pollack, C., Merkley, R., Ansari, D., & Price, G. R. (2020). The "inferior temporal numeral area" distinguishes numerals from other character categories during passive viewing: A representational similarity analysis. *NeuroImage*, *214*, 116716. https://doi.org/10.1016/j.neuroimage.2020.116716

Yeo, D. J., Wilkey, E. D., & Price, G. R. (2017). The search for the number form area: A functional neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, *78*, 145–160. https://doi.org/10.1016/j.neubiorev.2017.04.027

Zeki, S., & Marini, L. (1998). Three cortical stages of colour processing in the human brain. *Brain*, *121*(9), 1669–1685. https://doi.org/10.1093/brain/121.9.1669

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lee, J., Jung, M., Lustig, N., & Lee, J.-H. (2023). Neural representations of the perception of handwritten digits and visual objects from a convolutional neural network compared to humans. *Human Brain Mapping*, *44*(5), 2018–2038. https://doi.org/10.1002/hbm.26189