5-2018

Neural Encoding and Decoding with Deep Learning for Natural Vision

Haiguang Wen
Purdue University

NEURAL ENCODING AND DECODING WITH DEEP LEARNING FOR NATURAL VISION

by

Haiguang Wen

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Electrical & Computer Engineering

West Lafayette, Indiana

May 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Zhongming Liu, Ph.D., Chair
Weldon School of Biomedical Engineering
School of Electrical and Computer Engineering
Dr. Charles A. Bouman, Ph.D.
School of Electrical and Computer Engineering
Dr. Eugenio Culurciello, Ph.D.
Weldon School of Biomedical Engineering
Dr. Gregory S. Francis, Ph.D.
Department of Psychological Sciences
Dr. Stanley Chan, Ph.D.
School of Electrical and Computer Engineering

Approved by:
Dr. David Kozel
Head of the Graduate Program

For my loving family

ACKNOWLEDGMENTS

I was fortunate enough to join the Laboratory of Integrated Brain Imaging (LIBI) as the first Ph.D. student. I would like to express my deep gratitude to my major advisor, Dr. Zhongming Liu, for his exceptional guidance and support going far beyond what I could ever have imagined. It would not have been possible for me to enter the field of neuroscience without his guidance, encouragement and advice. As a strong engineer and scientist, he has taught me how to become a scientific researcher and what the right attitude should be in pursuing science. His broad and deep knowledge, insightful perspectives, and uncompromising attitude towards perfection, have deep and positive impacts on me. I sincerely thank him for helping me develop my scientific thinking, analytical skills, and communication skills. In addition, I am deeply grateful to my other committee members, Dr. Charles A. Bouman, Dr. Eugenio Culurciello, Dr. Gregory S. Francis, and Dr. Stanley Chan, for their critical comments and feedback at various stages of the PhD qualification process.

I would also like to thank all my colleagues in the LIBI for their valuable contributions to this dissertation work. It has been a pleasure to work with them in the past five years. I will always cherish our friendship. Thank you Kun-Han, Lauren, Junxing, Yizhen, Jiayue, Ranajay, Jung-Hoon, Kuan, Nishant, Steven, Jun Young, and Shao-Chin for their insightful discussions, critical feedbacks, and experiments help. I would also like to express my gratitude toward Dr. Wei Chen, Dr. Xiaohong Zhu, and Dr. Byeong-Yeul Lee for their constructive discussions and support in the functional magnetic resonance imaging (fMRI) experiments. I would also like to thank all my other friends who have given me support in various ways.

Finally, from the bottom of my heart, I would like to thank my parents, my grandparents, my loving brothers, and my uncle for their unconditional support and love, and for helping me become the person I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Term
CNN	Convolutional Neural Network
De-CNN	Deconvolutional Neural Network
DNN	Deep Neural Network
FDR	False Discovery Rate
FEF	Frontal Eye Field
FFA	Fusiform Face area
FM	Feature Map
fMRI	functional Magnetic Resonance Imaging
HRF	Haemodynamic Response Function
LIP	Lateral Intraparietal
LCH	Leacock-Chodorow Similarity
LO	Lateral Occipital
MT	Middle Temporal
OFA	Occipital Face Area
PC	Predictive Coding
PCA	Principal Component Analysis
PCN	Predictive Coding Network
PEF	Premotor Eye Field
PPA	Parahippocampal Place Area
pSTS	Posterior Superior Temporal Sulcus
RNN	Recurrent Neural Network
ROI	Region of Interest
SNR	Signal to Noise Ratio
TPJ	Temporo-Parietal Junction
VAE	Variational Autoencoder
VVC	Ventral Visual Complex

ABSTRACT

Author: Wen, Haiguang. PhD
Institution: Purdue University
Degree Received: May 2018
Title: Neural Encoding and Decoding with Deep Learning for Natural Vision.
Major Professor: Dr. Zhongming Liu, Ph.D.

The overarching objective of this work is to bridge neuroscience and artificial intelligence to ultimately build machines that learn, act, and think like humans. In the context of vision, the brain enables humans to readily make sense of the visual world, e.g. recognizing visual objects. Developing human-like machines requires understanding the working principles underlying the human vision. In this dissertation, I ask how the brain encodes and represents dynamic visual information from the outside world, whether brain activity can be directly decoded to reconstruct and categorize what a person is seeing, and whether neuroscience theory can be applied to artificial models to advance computer vision. To address these questions, I used deep neural networks (DNN) to establish encoding and decoding models for describing the relationships between the brain and the visual stimuli. Using the DNN, the encoding models were able to predict the functional magnetic resonance imaging (fMRI) responses throughout the visual cortex given video stimuli; the decoding models were able to reconstruct and categorize the visual stimuli based on fMRI activity. To further advance the DNN model, I have implemented a new bidirectional and recurrent neural network based on the predictive coding theory. As a theory in neuroscience, predictive coding explains the interaction among feedforward, feedback, and recurrent connections. The results showed that this brain-inspired model significantly outperforms feedforward-only DNNs in object recognition. These studies have positive impact on understanding the neural computations under human vision and improving computer vision with the knowledge from neuroscience.

1. INTRODUCTION

1.1 Defining the Problem

Recent progress in artificial intelligence and computational neuroscience converges to a new strategy to further advance both areas through their positive synergy [1-3]. In the context of natural vision, brain-inspired artificial models, e.g. the deep neural network (DNN), have achieved impressive state-of-the-art performance in understanding complex images and videos [4-6]. Comparing such models against the human visual system under natural vision has also led to in-depth understanding of how the brain represents visual information [7-10]. As such, computational neuroscience advances artificial intelligence, and vice versa. Explicitly linking the mechanisms of visual processing between biological brains and artificial models is expected to accelerate progress in both fields.

The overarching objective of this dissertation is to bridge neuroscience and artificial intelligence for ultimately building human-like machine vision. The human, as the best example of intelligence known, has unsurpassed ability for perceiving, processing and understanding complex and dynamic visual stimuli from the outside world. While current artificial intelligence benefits from gaining inspiration from neuroscience knowledge, there is still a long way to go before we fully understand biological brains. In vision, two major unresolved questions are 1) how the human brain represents and organizes visual information [11], and 2) whether brain activity can be decoded in real-time to reconstruct and interpret what a person is seeing [12]. Addressing these questions requires not only measurements of brain activity but also computational models with built-in hypotheses about neural computation and learning. So far, the brain-inspired deep neural networks have become the best computational models for processing visual information in natural images or videos [4]. Therefore, in the aspect of ***computational neuroscience***, I used the brain-inspired DNNs to model, predict, and decode brain activity during dynamic natural vision. In addition, though the current DNNs achieve state-of-the-art performances in some computer vision tasks, they are still far from the biological brain. To further advance the artificial models, it requires the models to be more brain-inspired. In the aspect of ***deep learning***, I developed a new brain-inspired recurrent neural network based on the predictive coding (PC) theory [13-16]. As a

theory in neuroscience, predictive coding explains the interaction among feedforward, feedback, and recurrent connections, which is essential to the network basis of natural vision.

1.2 Computational Neuroscience

For centuries, philosophers and scientists have been trying to speculate, observe, understand, and decipher the workings of the brain that enables humans to perceive and explore visual surroundings. Understanding the human visual system requires computational models with built-in hypotheses about neural computation and learning [2]. Models that truly reflect the brain's working in natural vision should be able to explain brain activity given any visual input (encoding) [12], and decode brain activity to infer visual input (decoding) [12]. Therefore, evaluating the model's encoding and decoding performance serves to test and compare hypotheses about how the brain learns and organizes visual representations [7].

Concerning the neural encoding and decoding, conventional neuroscience studies use artificial patterns or static pictures to identify neural representations of isolated visual elements or categories [12, 17, 18]. However, such strategies are too narrowly focused to reveal the computation underlying natural vision, which is highly dynamic, complex and diverse. What is needed is an alternative strategy that embraces the complexity of vision to uncover and decode the visual representations of neural activity. To date, deep learning provides the most comprehensive computational models to encode hierarchically organized features from natural pictures or videos [4]. Computer-vision systems based on such models have emulated or even surpassed human performance in image recognition and segmentation [6, 19-21]. In particular, deep convolutional neural networks (CNNs) are built and trained with similar organizational principles as the feedforward visual-cortical network [2, 3]. Therefore, I developed and used deep-learning models to study the neural encoding and decoding for natural vision.

In Chapter 2, I used a pretrained CNN driven for object recognition to establish 1) encoding models that predict the fMRI responses in the visual cortex given video stimuli and 2) decoding models that reconstruct and categorize the video stimuli given the fMRI activities [7]. CNN has been shown to be able to explain cortical responses to static pictures at ventral-stream areas [8-10]. Here, we further showed that such CNN could reliably predict fMRI responses from humans watching natural movies, despite its lack of any mechanism to account for temporal dynamics or feedback processing. For training and testing the encoding and decoding models, I acquired 44.8

hours of fMRI data from 3 human subjects when watching ~9,300 different video clips, including diverse objects, scenes and actions. This dataset was independent of, and had a much larger sample size and broader coverage than, those in prior studies [8-10, 22]. Through the encoding models, the CNN-predicted areas covered not only the ventral stream, but also the dorsal stream, albeit to a lesser degree; single-voxel response was visualized as the specific pixel pattern that drove the response, revealing the distinct representation of individual cortical location; cortical activation was synthesized from natural images with high-throughput to map category representation, contrast, and selectivity. Through the decoding models, the decoders supported direct visual reconstruction and semantic categorization of natural movies from the fMRI responses. The decoding models are efficient since it does not require comprehensive searching from large candidate stimuli given the observed activity pattern. This sets our method apart from multivariate pattern analysis [17, 18, 23] and encoding-model-based decoding [24-26].

In Chapter 3, I built and used DNN-based encoding models to study the visual representation and organization of natural visual objects [27]. The brain represents visual objects with topographic cortical patterns. To address how distributed visual representations enable object categorization, we established predictive encoding models based on a deep residual network [21], and trained them to predict cortical responses to natural movies. Using this predictive model, we mapped human cortical representations to 64,000 visual objects from 80 categories with high throughput and accuracy. Such representations covered both the ventral and dorsal pathways, reflected multiple levels of object features, and preserved semantic relationships between categories. In the entire visual cortex, object representations were organized into three clusters of categories: biological objects, non-biological objects, and background scenes. In a finer scale specific to each cluster, object representations revealed sub-clusters for further categorization. Such hierarchical clustering of category representations was mostly contributed by cortical representations of object features from middle to high levels. In summary, this study demonstrates a useful computational strategy to characterize the cortical organization and representations of visual features for rapid categorization.

In Chapter 4, I developed new methods for transferring and generalizing deep-learning-based encoding models across subjects [28]. Recent studies have shown the value of using deep learning models for mapping and characterizing how the brain represents and organizes information for natural vision [7-10, 22, 29, 30]. However, training the encoding models requires

measuring cortical responses to large and diverse sets of natural visual stimuli from single subjects. This requirement limits prior studies to few subjects, making it difficult to generalize findings across subjects or for a population. Here, I developed new methods to transfer and generalize encoding models across subjects. To train encoding models specific to a target subject, the models trained for other subjects were used as the prior models and were refined efficiently using Bayesian inference with a limited amount of data from the target subject. To train encoding models for a population, the models were progressively trained and updated with incremental data from different subjects. Results demonstrate that the proposed methods provide an efficient and effective strategy to establish both subject-specific and population-wide predictive models of cortical representations of high-dimensional and hierarchical visual features.

These studies have shown the unique value of using deep-learning models and video-fMRI dataset to map the hierarchical representation in the visual cortex, the cortical representation of object categories and the hierarchical distribution of process memory. As such, it provides an all-in-one strategy for mapping and characterizing various functional and computational aspects of human vision.

1.3 Deep Learning

Inspired by biological neural networks, convolutional neural networks have recently provided superior performance in image classification [4-6, 19-21], face recognition [31], scene parsing [32], object segmentation [33], to name a few. Deep neural networks have showed tremendous performance in very hard vision tasks, such as the ImageNet competition [34], where DNN are now practically the most successful algorithm used [6, 19-21]. Recent studies by us and others have also shown that deep CNNs can predict cascaded cortical processes underlying object perception [7-10, 22, 30].

Despite various ways of architectural reconfiguration, these DNNs all scale up from the same principle of computation: extracting image features by a feedforward pass through stacks of convolutional layers. However, the brain contains feedforward, recurrent and feedback connections, and their complex interactions give rise to visual perception, attention, and action. To mitigate this limitation, we have developed a recurrent neural network by adding recurrent connections to CNN. The recurrent model performed better in action recognition, better explained the brain responses to natural videos, and revealed the hierarchical distribution of process memory

[29]. In addition, our recent development on variational autoencoder (VAE) with feedback connections attempts to model the generative processes in the visual cortex [35]. However, there is to date no established model to fully explain dynamic interactions among feedforward, feedback and recurrent connections, which is essential to the network basis of natural vision.

In Chapter 5, to further advance the artificial model so that it becomes more brain-like, I have implemented a new bidirectional and recurrent neural network based on the predictive coding theory[13, 14, 16, 36-38], called the predictive coding network (PCN) [39]. As a theory in neuroscience, the predictive coding explains the interaction among feedforward, feedback and recurrent connections, supported by a number of neuroscience studies [15, 40-42]. Specifically, the feedback connections convey the top-down prediction of the representation at the lower level, while the feedforward connections propagate the residual error between the top-down prediction and the actual activity to the level above. Unlike CNN, RNN, or VAE, the predictive coding network includes feedforward, feedback, and recurrent connections, and accounts for their dynamic interactions given naturalistic visual inputs. The results showed that such brain-inspired model significantly outperforms the CNN in object recognition. Our development on the brain-inspired PCN sheds light on the use of artificial network in modeling the complex dynamic process. This dissertation research provides unique video-fMRI dataset and novel neural coding methods and models to identify the common architectural, computational and learning principles that support both computer vision and human vision. The dataset, models, and codes are publically available to facilitate research in neuroscience, computer science or other communities.

2. DEEP NEURAL NETWORK PREDICTS AND DECODES THE CORTICAL REPRESENTATION OF DYNAMIC VISUAL STIMULI

*Modified and formatted for dissertation from the article published in *Cerebral Cortex* [7]

2.1 Introduction

For centuries, philosophers and scientists have been trying to speculate, observe, understand, and decipher the workings of the brain that enables humans to perceive and explore visual surroundings. Here, we ask how the brain represents dynamic visual information from the outside world, and whether brain activity can be directly decoded to reconstruct and categorize what a person is seeing. These questions, concerning neural encoding and decoding [12], have been mostly addressed with static or artificial stimuli [17, 18]. Such strategies are, however, too narrowly focused to reveal the computation underlying natural vision. What is needed is an alternative strategy that embraces the complexity of vision to uncover and decode the visual representations of distributed cortical activity.

Despite its diversity and complexity, the visual world is composed of a large number of visual features [4, 43, 44]. These features span many levels of abstraction, such as orientation and color in the low level, shapes and textures in the middle levels, and objects and actions in the high level. To date, deep learning provides the most comprehensive computational models to encode and extract hierarchically organized features from arbitrary natural pictures or videos [4]. Computer-vision systems based on such models have emulated or even surpassed human performance in image recognition and segmentation [19, 21, 45]. In particular, deep convolutional neural networks (CNN) are built and trained with similar organizational and coding principles as the feedforward visual cortical network [2, 46]. Recent studies have shown that the CNN could partially explain the brain's responses to [8, 9, 30] and representations of [10, 22] natural picture stimuli. However, it remains unclear whether and to what extent the CNN may explain and decode brain responses to natural video stimuli. Although dynamic natural vision involves feedforward, recurrent, and feedback connections [47], the CNN only models feedforward processing and operates on instantaneous input, without any account for recurrent or feedback network interactions [13, 48].

To address these questions, we acquired 11.5 hours of fMRI data from each of three human subjects watching 972 different video clips, including diverse scenes and actions. This dataset was independent of, and had a larger sample size and broader coverage than, those in prior studies [8-10, 22, 30]. This allowed us to confirm, generalize, and extend the use of the CNN in predicting and decoding cortical activity along both ventral and dorsal streams in a dynamic viewing condition. Specifically, we trained and tested encoding and decoding models, with distinct data, for describing the relationships between the brain and the CNN, implemented by [19]. With the CNN, the encoding models were used to predict and visualize fMRI responses at individual cortical voxels given the movie stimuli; the decoding models were used to reconstruct and categorize the visual stimuli based on fMRI activity, as shown in Fig. 2.1. The major findings were

1) a CNN driven for image recognition explained significant variance of fMRI responses to complex movie stimuli for nearly the entire visual cortex including its ventral and dorsal streams, albeit to a lesser degree for the dorsal stream;
2) the CNN-based voxel-wise encoding models visualized different single-voxel representations, and revealed category representation and selectivity;
3) the CNN supported direct visual reconstruction of natural movies, highlighting foreground objects with blurry details and missing colors;
4) the CNN also supported direct semantic categorization, utilizing the semantic space embedded in the CNN.

2.2 Methods and Materials

2.2.1 Subjects and experiments

Three healthy volunteers (female, age: 22-25; normal vision) participated in the study, with informed written consent obtained from every subject according to the research protocol approved by the Institutional Review Board at Purdue University. Each subject was instructed to watch a series of natural color video clips ($20.3^o \times 20.3^o$) while fixating at a central fixation cross ($0.8^o \times 0.8^o$). In total, 374 video clips (continuous with a frame rate of 30 frames per second) were included in a 2.4-hour training movie, randomly split into 18 8-min segments; 598 different video clips were included in a 40-min testing movie, randomly split into five 8-min segments. The video clips in the testing movie were different from those in the training movie. All video clips were chosen from

Videoblocks (https://www.videoblocks.com) and YouTube (https://www.youtube.com) to be diverse yet representative of real-life visual experiences. For example, individual video clips showed people in action, moving animals, nature scenes, outdoor or indoor scenes etc. Each subject watched the training movie twice and the testing movie ten times through experiments in different days. Each experiment included multiple sessions of 8min and 24s long. During each session, an 8-min single movie segment was presented; before the movie presentation, the first movie frame was displayed as a static picture for 12 seconds; after the movie, the last movie frame was also displayed as a static picture for 12 seconds. The order of the movie segments was randomized and counter-balanced. Using Psychophysics Toolbox 3 (http://psychtoolbox.org), the visual stimuli were delivered through a goggle system (NordicNeuroLab NNL Visual System) with 800×600 display resolution.

2.2.2 Data acquisition and preprocessing

T_1 and T_2-weighted MRI and fMRI data were acquired in a 3 tesla MRI system (Signa HDx, General Electric Healthcare, Milwaukee) with a 16-channel receive-only phase-array surface coil (NOVA Medical, Wilmington). The fMRI data were acquired at 3.5 mm isotropic spatial resolution and 2 s temporal resolution by using a single-shot, gradient-recalled echo-planar imaging sequence (38 interleaved axial slices with 3.5mm thickness and $3.5\times3.5mm^2$ in-plane resolution, TR=2000ms, TE=35ms, flip angle=78°, field of view=$22\times22cm^2$). The fMRI data were preprocessed and then transformed onto the individual subjects' cortical surfaces, which were co-registered across subjects onto a cortical surface template based on their patterns of myelin density and cortical folding. The preprocessing and registration were accomplished with high accuracy by using the processing pipeline for the Human Connectome Project [49]. When training and testing the encoding and decoding models (as described later), the cortical fMRI signals were averaged over multiple repetitions: two repetitions for the training movie, and 10 repetitions for the testing movie. The two repetitions of the training movie allowed us to evaluate intra-subject reproducibility in the fMRI signal as a way to map the regions "activated" by natural movie stimuli. The ten repetitions of the testing movie allowed us to obtain the movie-evoked responses with high signal to noise ratios (SNR), as spontaneous activity or noise unrelated to visual stimuli were effectively removed by averaging over this relatively large number of repetitions. The ten repetitions of the testing movie also allowed us to estimate the upper bound (or "noise ceiling"),

by which an encoding model could predict the fMRI signal during the testing movie. Although more repetitions of the training movie would also help to increase the SNR of the training data, it was not done because the training movie was too long to repeat by the same times as the testing movie.

2.2.3 Convolutional neural network (CNN)

We used a deep CNN (a specific implementation referred as the "AlexNet") to extract hierarchical visual features from the movie stimuli. The model had been pre-trained to achieve the best-performing object recognition in Large Scale Visual Recognition Challenge 2012 [19]. Briefly, this CNN included eight layers of computational units stacked into a hierarchical architecture: the first five were convolutional layers, and the last three layers were fully connected for image-object classification. The image input was fed into the first layer; the output from one layer served as the input to its next layer. Each convolutional layer contained a large number of units and a set of filters (or kernels) that extracted filtered outputs from all locations from its input through a rectified linear function. Layer 1 through 5 consisted of 96, 256, 384, 384, and 256 kernels, respectively. Max-pooling was implemented between layer 1 and layer 2, between layer 2 and layer 3, and between layer 5 and layer 6. For classification, layer 6 and 7 were fully connected networks; layer 8 used a softmax function to output a vector of probabilities, by which an input image was classified into individual categories. The numbers of units in layer 6 to 8 were 4096, 4096, and 1000, respectively.

Note that the 2nd highest layer in the CNN (i.e. the 7th layer) effectively defined a semantic space to support the categorization at the output layer. In other words, the semantic information about the input image was represented by a (4096-dimensional) vector in this semantic space. In the original AlexNet, this semantic space was used to classify ~1.3 million natural pictures into 1,000 fine-grained categories [19]. Thus, it was generalizable and inclusive enough to also represent the semantics in our training and testing movies, and to support more coarsely defined categorization. Indeed, new classifiers could be built for image classification into new categories based on the generic representations in this same semantic space, as shown elsewhere for transfer learning [50].

Many of the 1,000 categories in the original AlexNet were not readily applicable to our training or testing movies. Thus, we reduced the number of categories to 15 for mapping

categorical representations and decoding object categories from fMRI. The new categories were coarser and labeled as *indoor*, *outdoor*, *people*, *face*, *bird*, *insect*, *water animal*, *land animal*, *flower*, *fruit*, *natural scene*, *car*, *airplane*, *ship*, and *exercise*. These categories covered the common content in both the training and testing movies. With the redefined output layer, we trained a new softmax classifier for the CNN (i.e. between the 7th layer and the output layer), but kept all lower layers unchanged. We used ~20,500 human-labeled images to train the classifier while testing it with a different set of ~3,500 labeled images. The training and testing images were all randomly and evenly sampled from the aforementioned 15 categories in ImageNet, followed by visual inspection to replace mis-labeled images.

In the softmax classifier (a multinomial logistic regression model), the input was the semantic representation, $\boldsymbol{y}$, from the 7th layer in the CNN, and the output was the normalized probabilities, $\boldsymbol{q}$, by which the image was classified into individual categories. The softmax classifier was trained by using the mini-batch gradient descend to minimize the Kullback-Leibler (KL) divergence from the predicted probability, $\boldsymbol{q}$, to the ground truth, $\boldsymbol{p}$, in which the element corresponding to the labeled category was set to one and others were zeros. The KL divergence indicated the amount of information lost when the predicted probability, $\boldsymbol{q}$, was used to approximate $\boldsymbol{p}$. The predicted probability was expressed as $\boldsymbol{q} = \frac{\exp(\boldsymbol{y}\mathbf{W}+\boldsymbol{b})}{\sum \exp(\boldsymbol{y}\mathbf{W}+\boldsymbol{b})}$, parameterized with $\mathbf{W}$ and $\boldsymbol{b}$. The objective function that was minimized for training the classifier was expressed as below.

$$D_{KL}(\boldsymbol{p} \,||\, \boldsymbol{q}) = H(\boldsymbol{p}, \boldsymbol{q}) - H(\boldsymbol{p}) = -\langle \boldsymbol{p}, \log \boldsymbol{q} \rangle + \langle \boldsymbol{p}, \log \boldsymbol{p} \rangle \qquad (1)$$

where $\mathrm{H}(\mathbf{p})$ was the entropy of $\mathbf{p}$, and $\mathrm{H}(\mathbf{p}, \mathbf{q})$ was the cross-entropy of $\mathbf{p}$ and $\mathbf{q}$, and $\langle \cdot \rangle$ stands for inner product. The objective function was minimized with L2-norm regularization whose parameter was determined by cross-validation. 3075 validation images (15% of the training images) were uniformly and randomly selected from each of the 15 categories. When training the model, the batch size was 128 samples per batch, the learning rate was initially 10^{-3} reduced by 10^{-6} every iteration. After training with 100 epochs, the classifier achieved a top-1 error of 13.16% with the images in the testing set.

Once trained, the CNN could be used for feature extraction and image recognition by a simple feedforward pass of an input image. Specifically, passing a natural image into the CNN resulted in an activation value at each unit. Passing every frame of a movie resulted in an activation time series from each unit, representing the fluctuating representation of a specific feature in the

movie. Within a single layer, the units that shared the same kernel collectively output a feature map given every movie frame. Herein we refer to the output from each layer as the output of the rectified linear function before max-pooling (if any).

2.2.4 Deconvolutional neural network (De-CNN)

While the CNN implemented a series of cascaded "bottom-up" transformations that extracted nonlinear features from an input image, we also used the De-CNN to approximately reverse the operations in the CNN, for a series of "top-down" projections as described in detail elsewhere [43]. Specifically, the outputs of one or multiple units could be unpooled, rectified, and filtered onto its lower layer, until reaching the input pixel space. The unpooling step was only applied to the layers that implemented max-pooling in the CNN. Since the max-pooling was non-invertible, the unpooling was an approximation while the locations of the maxima within each pooling region were recorded and used as a set of switch variables. Rectification was performed as point-wise rectified linear thresholding by setting the negative units to 0. The filtering step was done by applying the transposed version of the kernels in the CNN to the rectified activations from the immediate higher layer, to approximate the inversion of the bottom-up filtering. In the De-CNN, rectification and filtering were independent of the input, whereas the unpooling step was dependent on the input. Through the De-CNN, the feature representations at a specific layer could yield a reconstruction of the input image [43]. This was utilized for reconstructing the visual input based on the 1^{st}-layer feature representations estimated from fMRI data. Such reconstruction is unbiased by the input image, since the De-CNN did not perform unpooling from the 1^{st} layer to the pixel space.

2.2.5 Mapping cortical activations with natural movie stimuli

Each segment of the training movie was presented twice to each subject. This allowed us to map cortical locations activated by natural movie stimuli, by computing the intra-subject reproducibility in voxel time series [51, 52]. For each voxel and each segment of the training movie, the intra-subject reproducibility was computed as the correlation of the fMRI signal when the subject watched the same movie segment for the first time and for the second time. After converting the correlation coefficients to z scores by using the Fisher's z-transformation, the voxel-wise z scores were averaged across all 18 segments of the training movie. Statistical significance

was evaluated by using one-sample t-test ($p<0.01$, DOF=17, Bonferroni correction for the number of cortical voxels), revealing the cortical regions activated by the training movie. Then, the intra-subject reproducibility maps were averaged across the three subjects. The averaged activation map was used to create a cortical mask that covered all significantly activated locations. To be more generalizable to other subjects or stimuli, we slightly expanded the mask. The final mask contained 10,214 voxels in the visual cortex, approximately 17.2% of the whole cortical surface.

2.2.6 Bivariate analysis to relate CNN units to brain voxels

We compared the outputs of CNN units to the fMRI signals at cortical voxels during the training movie, by evaluating the correlation between every unit and every voxel. Before this bivariate correlation analysis, the single unit activity in the CNN was log-transformed and convolved with a canonical hemodynamic response function (HRF) with the positive peak at 4s. Such preprocessing was to account for the difference in distribution, timing, and sampling between the unit activity and the fMRI signal. The unit activity was non-negative and sparse; after log-transformation (i.e. $\log(y + 0.01)$ where y indicated the unit activity), it followed a distribution similar to that of the fMRI signal. The HRF accounted for the temporal delay and smoothing due to neurovascular coupling. Here, we preferred a pre-defined HRF to a model estimated from the fMRI data itself. While the latter was data-driven and used in previous studies [26, 53], it might cause over-fitting. A pre-defined HRF was suited for more conservative estimation of the bivariate (unit-to-voxel) relationships. Lastly, the HRF-convolved unit activity was down-sampled to match the sampling rate of fMRI. With such preprocessing, the bivariate correlation analysis was used to map the retinotopic, hierarchical, and categorical representations during natural movie stimuli, as described subsequently.

Retinotopic mapping. In the first layer of the CNN, individual units extracted features (e.g. orientation-specific edge) from different local (11-by-11 pixels) patches in the input image. We computed the correlation between the fMRI signal at each cortical location and the activation time series of every unit in the first layer of the CNN during the training movie. For a given cortical location, such correlations formed a 3-D array: two dimensions corresponding to the horizontal and vertical coordinates in the visual field, and the third dimension corresponding to 96 different local features (see Fig. 2.7 c). As such, this array represented the simultaneous tuning of the fMRI response at each voxel by retinotopy, orientation, color, contrast, spatial frequency etc. We reduced

the 3-D correlation array into a 2-D correlation matrix by taking the maximal correlation across different visual features. As such, the resulting correlation matrix depended only on retinotopy, and revealed the population receptive field (pRF) of the given voxel. The pRF center was determined as the centroid of the top 20 locations with the highest correlation values, and its polar angle and eccentricity were further measured with respect to the central fixation point. Repeating this procedure for every cortical location gave rise to the putative retinotopic representation of the visual cortex. We compared this retinotopic representation obtained with natural visual stimuli to the visual-field maps obtained with the standard retinotopic mapping as previously reported elsewhere [54].

Hierarchical mapping. The feedforward visual processing passes through multiple cascaded stages in both the CNN and the visual cortex. In line with previous studies [8-10, 22, 30, 53, 55, 56], we explored the correspondence between individual layers in the CNN and individual cortical regions underlying different stages of visual processing. For this purpose, we computed the correlations between the fMRI signal at each cortical location and the activation time series from each layer in the CNN, and extracted the maximal correlation. We interpreted this maximal correlation as a measure of how well a cortical location corresponded to a layer in the CNN. For each cortical location, we identified the best corresponding layer and assigned its layer index to this location; the assigned layer index indicated the processing stage this location belonged to. The cortical distribution of the layer-index assignment provided a map of the feedforward hierarchical organization of the visual system.

Mapping representations of object categories. To explore the correspondence between the high-level visual areas and the object categories encoded by the output layer of the CNN, we examined the cortical fMRI correlates to the 15 categories output from the CNN. Here, we initially focused on the "face" because face recognition was known to involve specific visual areas, such as the fusiform face area (FFA) [57, 58]. We computed the correlation between the activation time series of the face-labeled unit (the unit labeled as "face" in the output layer of the CNN) and the fMRI signal at every cortical location, in response to each segment of the training movie. The correlation was then averaged across segments and subjects. The significance of the average correlation was assessed using a block permutation test [59] in consideration of the auto-correlation in the fMRI signal. Specifically, the time series was divided into 50-sec blocks of adjacent 25 volumes (TR=2s). The block size was chosen to be long enough to account for the

autocorrelation of fMRI and to ensure a sufficient number of permutations to generate the null distribution. During each permutation step, the "face" time series underwent a random shift (i.e. removing a random number of samples from the beginning and adding them to the end) and then the time-shifted signal was divided into blocks, and permuted by blocks. For a total of 100,000 times of permutations, the correlations between the fMRI signal and the permuted "face" time series was calculated. This procedure resulted in a realistic null distribution, against which the p value of the correlation (without permutation) was calculated with Bonferroni correction by the number of voxels. The significantly correlated voxels ($p<0.01$) were displaced to reveal cortical regions responsible for the visual processing of human faces. The same strategy was also applied to the mapping of other categories.

2.2.7 Voxel-wise encoding models

Furthermore, we attempted to establish the CNN-based predictive models for the fMRI response to natural movie stimuli. Such models were defined separately for each voxel, namely voxel-wise encoding models [12], through which the voxel response was predicted from a linear combination of the feature representations of the input movie. Conceptually similar encoding models were previously explored with low-level visual features [24, 26] or high-level semantic features [60, 61], and more recently with hierarchical features extracted by the CNN from static pictures [8, 30]. Here, we extended these prior studies to focus on natural movie stimuli while using principal component analysis (PCA) to reduce the huge dimension of the feature space attained with the CNN.

Specifically, PCA was applied to the feature representations obtained from each layer of the CNN during the training movie. Principal components were retained to keep 99% of the variance while spanning a much lower-dimensional feature space, in which the representations followed a similar distribution as did the fMRI signal. This dimension reduction mitigated the potential risk of overfitting with limited training data. In the reduce feature space, the feature time series were readily comparable with the fMRI signal without additional nonlinear (log) transformation.

Mathematically, let $\mathbf{Y}_{\mathrm{o}}^{l}$ be the output from all units in layer l of the CNN; it is an m-by-p matrix (m is the number of video frames in the training movie, and p is the number of units). The time series extracted by each unit was standardized (i.e. remove the mean and normalize the

variance). Let $\mathbf{B}^l$ be the principal basis of $\mathbf{Y}_o^l$; it is a *p*-by-*q* matrix (*q* is the number of components). Converting the feature representations from the unit-wise space to the component-wise space is expressed as below.

$$\mathbf{Y}_n^l = \mathbf{Y}_o^l \mathbf{B}^l \qquad (2)$$

where $\mathbf{Y}_n^l$ is the transformed feature representations in the dimension-reduced feature space spanned by unitary columns in the matrix, $\mathbf{B}^l$. The transpose of $\mathbf{B}^l$ also defined the transformation back to the original space.

Following the dimension reduction, the feature time series, $\mathbf{Y}_n^l$, were convolved with a HRF, and then down-sampled to match the sampling rate of fMRI. Hereafter, $\mathbf{Y}^l$ stands for the feature time series for layer l after convolution and down-sampling. These feature time series were used to predict the fMRI signal at each voxel through a linear regression model, elaborated as below.

Given a voxel v, the voxel response $\boldsymbol{x}_v$ was modeled as a linear combination of the feature time series, $\mathbf{Y}^l$, from the *l*-th layer in the CNN, as expressed in Eq. (3).

$$\boldsymbol{x}_v = \mathbf{Y}^l \boldsymbol{w}_v^l + b_v^l + \boldsymbol{\varepsilon} \qquad (3)$$

where, $\boldsymbol{w}_v^l$ is a *q*-by-1 vector of the regression coefficients; b_v^l is the bias term; $\boldsymbol{\varepsilon}$ is the error unexplained by the model. Least-squares estimation with L2-norm regularization, as Eq. (4), was used to estimate the regression coefficients based on the data during the training movie.

$$f(\boldsymbol{w}_v^l) = \|\boldsymbol{x}_v - \boldsymbol{Y}^l \boldsymbol{w}_v^l - b_v^l\|_2^2 + \lambda \|\boldsymbol{w}_v^l\|_2^2 \qquad (4)$$

Here, the L2 regularization was used to prevent the model from overfitting limited training data. The regularization parameter λ and the layer index *l* were both optimized through a nine-fold cross-validation. Briefly, the training data were equally split into nine subsets: eight for the model estimation, one for the model validation. The validation was repeated nine times such that each subset was used once for validation. The parameters (λ, *l*) were chosen to maximize the cross-validation accuracy. With the optimized parameters, we refitted the model using the entire training samples to yield the final estimation of the voxel-wise encoding model. The final encoding model set up a computational pathway from the visual input to the evoked fMRI response at each voxel via its most predictive layer in the CNN.

After training the encoding model, we tested the model's accuracy in predicting the fMRI response to all five segments of the testing movie, for which the model was not trained. For each voxel, the prediction accuracy was measured as the correlation between the measured fMRI response and the response predicted by the voxel-specific encoding model, averaged across the

segments of the testing movie. The significance of the correlation was assessed using a block permutation test [59], while considering the auto-correlation in the fMRI signal, similarly as the significance test for the unit-to-voxel correlation. Briefly, the predicted fMRI signal was randomly block-permuted in time for 100,000 times to generate an empirical null distribution, against which the prediction accuracy was evaluated for significance ($p<0.001$, Bonferroni correction by the number of voxels). The prediction accuracy was also evaluated for regions of interest (ROIs) defined with multi-modal cortical parcellation [62]. For the ROI analysis, the voxel-wise prediction accuracy was averaged within each ROI. The prediction accuracy was evaluated for each subject, and then compared and averaged across subjects.

The prediction accuracy was compared with an upper bound by which the fMRI signal was explainable by the visual stimuli, given the presence of noise or ongoing activity unrelated to the stimuli. This upper bound, defining the explainable variance for each voxel, depended on the signal to noise ratio of the evoked fMRI response. It was measured voxel by voxel based on the fMRI signals observed during repeated presentations of the testing movie. Specifically, 10 repetitions of the testing movie were divided by half. This two-half partition defined an (ideal) control model: the signal averaged within the first half was used to predict the signal averaged within the second half. Their correlation, as the upper bound of the prediction accuracy, was compared with the prediction accuracy obtained with the voxel-wise encoding model in predicting the same testing data. The difference between their prediction accuracies (z-score) was assessed by paired t-test ($p<0.01$) across all possible two-half partitions and all testing movie segments. For those significant voxels, we then calculated the percentage of the explainable variance that was not explained by the encoding model. Specifically, let V_c be the potentially explainable variance; let V_e be the variance explained by the encoding model; so, $(V_c - V_e)/V_c$ measures the degree by which the encoding falls short in explaining the stimulus-evoked response [63].

2.2.8 Predicting cortical responses to images and categories

After testing their ability to predict cortical responses to unseen stimuli, we further used the encoding models to predict voxel-wise cortical responses to arbitrary pictures. Specifically, 15,000 images were uniformly and randomly sampled from 15 categories in ImageNet (i.e. *face, people, exercise, bird, land-animal, water-animal, insect, flower, fruit, car, airplane, ship, natural scene, outdoor, indoor*). None of these sampled images were used to train the CNN, or included

in the training or testing movies. For each sampled image, the response at each voxel was predicted by using the voxel-specific encoding model. The voxel's responses to individual images formed a response profile, indicative of its selectivity to single images.

To quantify how a voxel selectively responded to images from a given category (e.g. face), the voxel's response profile was sorted in a descending order of its response to every image. Since each category contained 1,000 exemplars, the percentage of the top-1000 images belonging to one category was calculated as an index of the voxel's categorical selectivity. This selectivity index was tested for significance using a binomial test against a null hypothesis that the top 1,000 images were uniformly random across individual categories. This analysis was tested specifically for voxels in the fusiform face area (FFA).

For each voxel, its categorical representation was obtained by averaging single-image responses within categories. The representational difference between inanimate vs. animate categories was assessed, with former including *flower, fruit, car, airplane, ship, natural scene, outdoor, indoor*, and the latter including *face, people, exercise, bird, land-animal, water-animal, insect*. The significance of this difference was assessed with two-sample t-test with Bonferroni correction by the number of voxels.

2.2.9 Visualizing single-voxel representations

The voxel-wise encoding models set up a computational path to relate any visual input to the evoked fMRI response at each voxel. It inspired and allowed us to reveal which part of the visual input specifically accounted for the response at each voxel, or to visualize the voxel's representation of the input. Note that the visualization was targeted to each voxel, as opposed to a layer or unit in the CNN, as in [8]. This distinction was important because voxels with activity predictable by the same layer in the CNN, may bear highly or entirely different representations.

Let us denote the visual input as $\mathbf{I}$. The response $\boldsymbol{x}_v$ at a voxel v was modeled as $\boldsymbol{x}_v = \mathrm{E}_v(\mathbf{I})$ (E_v is the voxel's encoding model). The voxel's visualized representation was an optimal gradient pattern given the visual input $\mathbf{I}$ that reflected the pixel-wise influence in driving the voxel's response. This optimization included two steps, combining the visualization methods based on masking [64, 65] and gradient [20, 66-68].

Firstly, the algorithm searched for an optimal binary mask, $\mathbf{M}_o$, such that the masked visual input gave rise to the maximal response at the target voxel, as Eq. (5).

$$\mathbf{M}_o = \arg\max_{\mathrm{M}}\{\mathrm{E}_v(\mathbf{I} \circ \mathbf{M})\} \qquad (5)$$

where the mask was a 2-D matrix with the same width and height as the visual input **I**, and ∘ stands for the Hadamard product, meaning that the same masking was applied to the red, green, and blue channels respectively. Since the encoding model was highly nonlinear and not convex, random optimization [69] was used. A binary continuous mask (i.e. the pixel weights were either 1 or 0) was randomly and iteratively generated. For each iteration, a random pixel pattern was generated with each pixel's intensity sampled from a normal distribution; this random pattern was spatially smoothed with a Gaussian spatial-smoothing kernel (three times of the kernel size of 1st layer CNN units); the smoothed pattern was thresholded by setting one fourth pixels to 1 and others 0. Then, the model-predicted response was computed given the masked input. The iteration was stopped when the maximal model-predicted response (over all iterations) converged or reach 100 iterations. The optimal mask was the one with the maximal response across iterations.

After the mask was optimized, the input from the masked region, $\mathbf{I}_o = \mathbf{I} \circ \mathbf{M}_o$, was supplied to the voxel-wise encoding model. The gradient of the model's output was computed with respect to the intensity at every pixel in the masked input, as expressed by Eq. (6). This gradient pattern described the relative influence of every pixel in driving the voxel response. Only positive gradients, which indicated the amount of influence in increasing the voxel response, were back-propagated and kept, as in [68].

$$\mathbf{G}_v(\mathbf{I}_o) = \nabla \mathrm{E}_v(\mathbf{I})|_{\mathbf{I}=\mathbf{I}_o} \qquad (6)$$

For the visualization to be more robust, the above two steps were repeated 100 times. The weighted average of the visualizations across all repeats was obtained with the weight proportional to the response given the masked input for each repeat (indexed with i), as Eq. (7). Consequently, the averaged gradient pattern was taken as the visualized representation of the visual input at the given voxel.

$$\mathbf{G}_v(\mathbf{I}_o) = \frac{1}{100}\sum_{i=1}^{100} \mathbf{G}_v^i(\mathbf{I}_o)\mathrm{E}_v^i(\mathbf{I}_o) \qquad (7)$$

This visualization method was applied to the fMRI signals during one segment of the testing movie. To explore and compare the visualized representations at different cortical locations, example voxels were chosen from several cortical regions across different levels, including V2, V4, MT, LO, FFA and PPA. Within each of these regions, we chose the voxel with the highest average prediction accuracy during the other four segments of the testing movie. The single-voxel

representations were visualized only at time points where peak responses occurred at one or multiple of the selected voxels.

2.2.10 Reconstructing natural movie stimuli

Opposite to voxel-wise encoding models that related visual input to fMRI signals, decoding models transformed fMRI signals to visual and semantic representations. The former was used to reconstruct the visual input, and the latter was used to uncover its semantics.

For the visual reconstruction, multivariate linear regression models were defined to take as input the fMRI signals from all voxels in the visual cortex, and to output the representation of every feature encoded by the 1st layer in the CNN. As such, the decoding models were feature-wise and multivariate. For each feature, the decoding model had multiple inputs and multiple outputs (i.e. representations of the given feature from all spatial locations in the visual input), and the times of fMRI acquisition defined the samples for the model's input and output. Eq. (8) describes the decoding model for each of 96 different visual features.

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \boldsymbol{\varepsilon} \qquad (8)$$

Here, $\mathbf{X}$ stands for the observed fMRI signals within the visual cortex. It is an m-by-(k+1) matrix, where m is the number of time points, k is the number of voxels; the last column of $\mathbf{X}$ is a constant vector with all elements equal to 1. $\mathbf{Y}$ stands for the log-transformed time-varying feature map. It is an m-by-p matrix, where m is the number of time points, and p is the number of units that encode the same local image feature (i.e. the convolutional kernel). $\mathbf{W}$ stands for the unknown weights, by which the fMRI signals are combined across voxels to predict the feature map. It is an (k+1)-by-p matrix with the last row being the bias component. $\boldsymbol{\varepsilon}$ is the error term.

To estimate the model, we optimized $\mathbf{W}$ to minimize the objective function below.

$$f(\mathbf{W}) = \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_2^2 + \lambda\|\mathbf{W}\|_1^1 \qquad (9)$$

where the first term is the sum of squares of the errors; the second term is the L1 regularization on $\mathbf{W}$ except for the bias component; λ is the hyperparameter balancing these two terms. Here, L1 regularization was used rather than L2 regularization, since the former favored sparsity as each visual feature in the 1st CNN layer was expected to be coded by a small set of voxels in the visual cortex [24, 70].

The model estimation was based on the data collected with the training movie. λ was determined by 20-fold cross-validation, similar to the procedures used for training the encoding

models. For training, we used stochastic gradient descent optimization with the batch size of 100 samples, i.e. only 100 fMRI volumes were utilized in each iteration of training. To address the overfitting problem, dropout technique [71] was used by randomly dropping 30% of voxels in every iteration, i.e. setting the dropped voxels to zeros. Dropout regularization was used to mitigate the co-linearity among voxels and counteract L1 regularization to avoid over-sparse weights. For the cross-validation, we evaluated for each of the 96 features, the validation accuracy defined as the correlation between the fMRI-estimated feature map and the CNN-extracted feature map. After sorting the different features in a descending order of the validation accuracy, we identified those features with relatively low cross-validation accuracy ($r < 0.24$), and excluded them when reconstructing the testing movie.

To test the trained decoding model, we applied it to the fMRI signals observed during one of the testing movies, according to Eq. (8) without the error term. To evaluate the performance of the decoding model, the fMRI-estimated feature maps were correlated with those extracted from the CNN given the testing movie. The correlation coefficient, averaged across different features, was used as a measure of the accuracy for visual reconstruction. To test the statistical significance of the reconstruction accuracy, a block permutation test was performed. Briefly, the estimated feature maps were randomly block-permuted in time [59] for 100,000 times to generate an empirical null distribution, against which the estimation accuracy was evaluated for significance ($p<0.01$), similar to the aforementioned statistical test for the voxel-wise encoding model.

To further reconstruct the testing movie from the fMRI-estimated feature maps, the feature maps were individually converted to the input pixel space using the De-CNN, and then were summed to generate the reconstruction of each movie frame. It is worth noting that the De-CNN did not perform unpooling from the 1st layer to the pixel space; so, the reconstruction was unbiased by the input, making the model generalizable for reconstruction of any unknown visual input. As a proof of concept, the visual inputs could be successfully reconstructed through De-CNN given the accurate (noiseless) feature maps [7].

2.2.11 Semantic categorization

In addition to visual reconstruction, the fMRI measurements were also decoded to deduce the semantics of each movie frame at the fMRI sampling times. The decoding model for semantic categorization included two steps: 1) converting the fMRI signals to the semantic representation

of the visual input in a generalizable semantic space, 2) converting the estimated semantic representation to the probabilities by which the visual input belonged to pre-defined and human-labeled categories.

In the first step, the semantic space was spanned by the outputs from the 7th CNN layer, which directly supported the image classification at the output layer. This semantic space was generalizable to not only novel images, but also novel categories which the CNN was not trained for [50]. As defined in Eq. (10), the decoding model used the fMRI signals to estimate the semantic representation, denoted as $\mathbf{Y}_s$ (m-by-q matrix, where q is the dimension of the dimension-reduced semantic space and m is the number of time points).

$$\mathbf{Y}_s = \mathbf{X}\mathbf{W}_s + \boldsymbol{\varepsilon} \quad (10)$$

where $\mathbf{X}$ stands for the observed fMRI signals within the visual cortex, and $\mathbf{W}_s$ was the regression coefficients, and $\boldsymbol{\varepsilon}$ was the error term. To train this decoding model, we used the data during the training movie and applied L2-regularization. The estimated dimension-reduced representation was then transformed back to the original space. The regularization parameter and q were determined by 9-fold cross validation based on the correlation between estimated representation and the ground truth.

In the second step, the semantic representation estimated in the first step was converted to a vector of normalized probabilities over categories. This step utilized the softmax classifier established when retraining the CNN for image classification into 15 labeled categories.

After estimating the decoding model with the training movie, we applied it to the data during one of the testing movies. It resulted in the decoded categorization probability for individual frames in the testing movie sampled every 2 seconds. The top 5 categories with the highest probabilities were identified, and their textual labels were displayed as the semantic descriptions of the reconstructed testing movie.

To evaluate the categorization accuracy, we used top-1 through top-3 prediction accuracies. Specifically, for any given movie frame, we ranked the object categories in a descending order of the fMRI-estimated probabilities. If the true category was the top 1 of the ranked categories, it was considered to be top-1 accurate. If the true category was in the top 2 of the ranked categories, it was considered to be top-2 accurate, so on and so forth. The percentage of the frames that were top-1/top-2/top-3 accurate was calculated to quantify the overall categorization accuracy, for which the significance was evaluated by a binomial test against the null hypothesis that the

categorization accuracy was equivalent to the chance level given random guesses. Note that the ground-truth categories for the testing movie was manually labeled by human observers, instead of the CNN's categorization of the testing movie.

2.2.12 Cross-subject encoding and decoding

To explore the feasibility of establishing encoding and decoding models generalizable to different subjects, we first evaluated the inter-subject reproducibility of the fMRI voxel response to the same movie stimuli. For each segment of the training movie, we calculated for each voxel the correlation of the fMRI signals between different subjects. The voxel-wise correlation coefficients were z-transformed and then averaged across all segments of the training movie. We assessed the significance of the reproducibility against zeros by using one-sample t-test with the degree of freedom as the total number of movie segments minus 1 (DOF=17, Bonferroni correction for the number of voxels, and $p<0.01$).

For inter-subject encoding, we used the encoding models trained with data from one subject to predict another subject's cortical fMRI responses to the testing movie. The accuracy of inter-subject encoding was evaluated in the same way as done for intra-subject encoding (i.e. training and testing encoding models with data from the same subject). For inter-subject decoding, we used the decoding models trained with one subject's data to decode another subject's fMRI activity for reconstructing and categorizing the testing movie. The performance of inter-subject decoding was evaluated in the same way as for intra-subject decoding (i.e. training and testing decoding models with data from the same subject).

2.3 Results

2.3.1 Functional alignment between CNN and visual cortex

For exploring and modeling the relationships between the CNN and the brain, we used 374 video clips to constitute a training movie, presented twice to each subject for fMRI acquisition. From the training movie, the CNN extracted visual features through hundreds of thousands of units, which were organized into eight layers to form a trainable bottom-up network architecture. That is, the output of one layer was the input to its next layer. After the CNN was trained for image categorization [19], each unit encoded a particular feature through its weighted connections to its lower layer, and its output reported the representation of the encoded feature in the input image.

The 1st layer extracted local features (e.g. orientation, color, contrast) from the input image; the 2nd through 7th layers extracted features with increasing nonlinearity, complexity, and abstraction; the highest layer reported the categorization probabilities [2, 4, 19].

The hierarchical architecture and computation in the CNN appeared similar to the feedforward processing in the visual cortex [2]. This motivated us to ask whether individual cortical locations were functionally similar to different units in the CNN given the training movie as the common input to both the brain and the CNN. To address this question, we first mapped the cortical activation with natural vision by evaluating the intra-subject reproducibility of fMRI activity when the subjects watched the training movie for the first vs. second time [51, 52]. The resulting cortical activation was widespread over the entire visual cortex (Fig. 2.2 a) for all subjects [7]. Then, we examined the relationship between the fMRI signal at every activated location and the output time series of every unit in the CNN. The latter indicated the time-varying representation of a particular feature in every frame of the training movie. The feature time series from each unit was log-transformed and convolved with the HRF, and then its correlation to each voxel's fMRI time series was calculated.

This bivariate correlation analysis was initially restricted to the 1st layer in the CNN. Since the 1st-layer units filtered the image patches with a fixed size at a variable location, their correlations with a voxel's fMRI signal revealed its population receptive field (pRF). The bottom insets in Fig. 2. 2.b. show the putative pRF of two example locations corresponding to peripheral and central visual fields. The retinotopic property was characterized by the polar angle and eccentricity of the center of every voxel's pRF [7], and mapped on the cortical surface (Fig. 2.2.b). The resulting retinotopic representations were consistent across subjects, and similar to the maps obtained with standard retinotopic mapping [54, 72]. The retinotopic organization reported here appeared more reasonable than the results obtained with a similar analysis approach but with natural picture stimuli [30], suggesting an advantage of using movie stimuli for retinotopic mapping than using static pictures. Beyond retinotopy, we did not observe any orientation-selective representations (i.e. orientation columns), most likely due to the low spatial resolution of the fMRI data.

Extending the above bivariate analysis beyond the 1st-layer of the CNN, different cortical regions were found to be preferentially correlated with distinct layers in the CNN (Fig. 2.2.c). The lower to higher level features encoded by the 1st through 8th layers in the CNN were gradually

mapped onto areas from the striate to extrastriate cortex along both ventral and dorsal streams (Fig. 2.2.c), consistently across subjects. These results agreed with findings from previous studies obtained with different analysis methods and static picture stimuli [8, 10, 22, 30, 53]. We extended these findings to further show that the CNN could map the hierarchical stages of feedforward processing underlying dynamic natural vision, with a rather simple and effective analysis method.

Furthermore, an investigation of the categorical features encoded in the CNN revealed a close relationship with the known properties of some high-order visual areas. For example, a unit labeled as "face" in the output layer of the CNN was significantly correlated with multiple cortical areas (Fig. 2.2.d, right), including the fusiform face area (FFA), the occipital face area (OFA), and the face-selective area in the posterior superior temporal sulcus (pSTS-FA), all of which have been shown to contribute to face processing [73]. Such correlations were also relatively stronger on the right hemisphere than on the left hemisphere, in line with the right hemispheric dominance observed in many face-specific functional localizer experiments [74]. In addition, the fMRI response at the FFA and the output of the 'face' unit both showed notable peaks coinciding with movie frames that included human faces (Fig. 2.2.d, left). These results exemplify the utility of mapping distributed neural-network representations of object categories automatically detected by the CNN. In this sense, it is more convenient than doing so by manually labeling movie frames, as in prior studies [44, 60]. Similar strategies were also used to reveal the network representations of 'indoor scenes', 'land animals', 'car', and 'bird' (Fig. 2.2.e).

Taken together, the above results suggest that the hierarchical layers in the CNN implement similar computational principles as cascaded visual areas along the brain's visual pathways. The CNN and the visual cortex not only share similar representations of some low-level visual features (e.g. retinotopy) and high-level semantic features (e.g. face), but also share similarly hierarchical representations of multiple intermediate levels of progressively abstract visual information (Fig. 2.2).

2.3.2 Neural encoding

Given the functional alignment between the human visual cortex and the CNN as demonstrated above and previously by others [8, 22, 30], we further asked whether the CNN could be used as a predictive model of the response at any cortical location given any natural visual input. In other words, we attempted to establish a voxel-wise encoding model by which the fMRI

response at each voxel was predicted from the output of the CNN. Specifically, for any given voxel, we optimized a linear regression model to combine the outputs of the units from a single layer in CNN to best predict the fMRI response during the training movie. We identified and used the principal components of the CNN outputs as the regressors to explain the fMRI voxel signal. Given the training movie, the output from each CNN layer could be largely explained by much fewer components. For the 1st through 8th layers, 99% of the variance in the outputs from 290400, 186624, 64896, 64896, 43264, 4096, 4096, 1000 units could be explained by 10189, 10074, 9901, 10155, 10695, 3103, 2804, 241 components, respectively. Despite dramatic dimension reduction especially for the lower layers, information loss was negligible (1%), and the reduced feature dimension largely mitigated overfitting when training the voxel-wise encoding model.

After training a separate encoding model for every voxel, we used the models to predict the fMRI responses to five 8-min testing movies. These testing movies included different video clips from those in the training movie, and thus unseen by the encoding models to ensure unbiased model evaluation. The prediction accuracy (r), measured as the correlation between the predicted and measured fMRI responses, was evaluated for every voxel. As shown in Fig. 2.3.a, the encoding models could predict cortical responses with reasonably high accuracies for nearly the entire visual cortex, much beyond the spatial extent predictable with low-level visual features [26] or high-level semantic features [60] alone. The model-predictable cortical areas shown in this study also covered a broader extent than was shown in prior studies using similar CNN-based feature models [8, 30]. The predictable areas even extended beyond the ventral visual stream, onto the dorsal visual stream, as well as areas in parietal, temporal, and frontal cortices (Fig. 2.3.a). These results suggest that object representations also exist in the dorsal visual stream, in line with prior studies [75, 76].

Regions of interest (ROI) were selected as example areas in various levels of visual hierarchy: V1, V2, V3, V4, lateral occipital (LO), middle temporal (MT), fusiform face area (FFA), parahippocampal place area (PPA), lateral intraparietal (LIP), temporo-parietal junction (TPJ), premotor eye field (PEF), and frontal eye field (FEF). The prediction accuracy, averaged within each ROI, was similar across subjects, and ranged from 0.4 to 0.6 across the ROIs within the visual cortex and from 0.25 to 0.3 outside the visual cortex (Fig. 2.3.b). These results suggest that the internal representations of the CNN explain cortical representations of low, middle, and high-level visual features to similar degrees. Different layers in the CNN contributed differentially to the

prediction at each ROI (Fig. 2.3.c). Also see Fig. 2.6.a for the comparison between the predicted and measured fMRI time series during the testing movie at individual voxels.

Although the CNN-based encoding models predicted partially but significantly the widespread fMRI responses during natural movie viewing, we further asked where and to what extent the models failed to fully predict the movie-evoked responses. Also note that the fMRI measurements contained noise and reflected in part spontaneous activity unrelated to the movie stimuli. In the presence of the noise, we defined a control model, in which the fMRI signal averaged over five repetitions of the testing movie was used to predict the fMRI signal averaged over the other five repetitions of the same movie. This control model served to define the explainable variance for the encoding model, or the ideal prediction accuracy (Fig. 2.4.a), against which the prediction accuracy of the encoding models (Fig. 2.4.b) was compared. Relative to the explainable variance, the CNN model tended to be more predictive of ventral visual areas (Fig. 2.4.c), which presumably sub-served the similar goal of object recognition as did the CNN. In contrast, the CNN model still fell relatively short in predicting the responses along the dorsal pathway (Fig. 2.4.c), likely because the CNN did not explicitly extract temporal features that are important for visual action [51].

2.3.3 Cortical representations of single-pictures or categories

The voxel-wise encoding models provided a fully computable pathway through which any arbitrary picture could be transformed to the stimulus-evoked fMRI response at any voxel in the visual cortex. As initially explored before [30], we conducted a high-throughput "virtual-fMRI" experiment with 15,000 images randomly and evenly sampled from 15 categories in ImageNet [34, 45]. These images were taken individually as input to the encoding model to predict their corresponding cortical fMRI responses. As a result, each voxel was assigned with a predicted response to every picture, and its response profile across individual pictures reported the voxel's functional representation [77]. For an initial proof of concept, we selected a single voxel that showed the highest prediction accuracy within FFA – an area for face recognition [57, 73, 74]. This voxel's response profile, sorted by the response level, showed strong face selectivity (Fig. 2.5.a). The top 1,000 pictures that generated the strongest responses at this voxel were mostly human faces (94.0%, 93.9%, and 91.9%) (Fig. 2.5.b). Such a response profile was not only limited to the selected voxel, but shared across a network including multiple areas from both hemispheres,

e.g. FFA, OFA, and pSTS-FA (Fig. 2.5c). It demonstrates the utility of the CNN-based encoding models for analyzing the categorical representations in voxel, regional, and network levels. Extending from this example, we further compared the categorical representation of every voxel, and generated a contrast map for the differential representations of animate vs. inanimate categories (Fig. 2.5d). We found that the lateral and inferior temporal cortex (including FFA) was relatively more selective to animate categories, whereas the parahippocampal cortex was more selective to inanimate categories (Fig. 2.5.d), in line with previous findings [78, 79].

2.3.4 Visualizing single-voxel representations given natural visual input

Not only could the voxel-wise encoding models predict how a voxel responded to different pictures or categories, such models were also expected to reveal how different voxels extract and process different visual information from the same visual input. To this end, we developed a method to visualize for each single voxel its representation given a known visual input. The method was to identify a pixel pattern from the visual input that accounted for the voxel response through the encoding model, revealing the voxel's representation of the input.

To visualize single-voxel representations, we selected six voxels from V2, V4, LO, MT, FFA and PPA (as shown in Fig. 2.6.a, left) as example cortical locations at different levels of visual hierarchy. For these voxels, the voxel-wise encoding models could well predict their individual responses to the testing movie (Fig. 2.6.a, right). At 20 time points when peak responses were observed at one or multiple of these voxels, the visualized representations shed light on their different functions (Fig. 2.6). It was readily notable that the visual representations of the V2 voxel were generally confined to a fixed part of the visual field, and showed pixel patterns with local details; the V4 voxel mostly extracted and processed information about foreground objects rather than from the background; the MT voxel selectively responded to the part of the movie frames that implied motion or action; the LO voxel represented either body parts or facial features; the FFA voxel responded selectively to human and animal faces, whereas the PPA voxel revealed representations of background, scenes, or houses. These visualizations offered intuitive illustration of different visual functions at different cortical locations, extending beyond their putative receptive-field size and location.

2.3.5 Neural decoding

While the CNN-based encoding models described the visual representations of individual voxels, it is the distributed patterns of cortical activity that gave rise to realistic visual and semantic experiences. To account for distributed neural coding, we sought to build a set of decoding models that combine individual voxel responses in a way to reconstruct the visual input to the eyes (visual reconstruction), and to deduce the visual percept in the mind (semantic categorization). Unlike previous studies [24, 26, 80-82], our strategy for decoding was to establish a computational path to directly transform fMRI activity patterns onto individual movie frames and their semantics captured at the fMRI sampling times.

2.3.6 Visual reconstruction

For visual reconstruction, we defined and trained a set of multivariate linear regression models to combine the fMRI signals across cortical voxels in an optimal way to match every feature map in the 1st CNN layer during the training movie. Such feature maps resulted from extracting various local features from every frame of the training movie (Fig. 2.7.a). By 20-fold cross-validation within the training data, the models tended to give more reliable estimates for 45 (out of 96) feature maps (Fig. 2.7.b), mostly related to features for detecting orientations and edges, whereas the estimates were less reliable for most color features (Fig. 2.7.c). In the testing phase, the trained models were used to convert distributed cortical responses generated by the testing movie to the estimated feature maps for the 1st-layer features. The reconstructed feature maps were found to be correlated with the actual feature maps directly extracted by the CNN ($r=0.30\pm0.04$). By using the De-CNN, every estimated feature map was transformed back to the pixel space, where they were combined to reconstruct the individual frames of the testing movie. Fig. 2.8 shows some examples of the movie frames reconstructed from fMRI vs. those actually presented. The reconstruction clearly captured the location, shape, and motion of salient objects, despite missing color. Perceptually less salient objects and the background were poorly reproduced in the reconstructed images. Such predominance of foreground objects is likely attributed to the effects of visual salience and attention on fMRI activity [83, 84]. Thus, the decoding in this study does not simply invert retinotopy [82] to reconstruct the original image, but tends to reconstruct the image parts relevant to visual perception. Miyawaki et al. previously used a similar computational strategy for direct reconstruction of simple pixel patterns, e.g. letters and shapes, with binary-

valued local image bases [85]. In contrast to the method in that study, the decoding method in this study utilized data-driven and biologically-relevant visual features to better account for natural image statistics [70, 86]. In addition, the decoding models, when trained and tested with natural movie stimuli, represented an apparently better account of cortical activity underlying natural vision, than the model trained with random images and tested for small-sized artificial stimuli [85].

2.3.7 Semantic categorization

To identify object categories from fMRI activity, we optimized a decoding model to estimate the category that each movie frame belonged to. Briefly, the decoding model included two parts: 1) a multivariate linear regression model that used the fMRI signals to estimate the semantic representation in the 7^{th} (i.e. the 2^{nd}-highest) CNN layer, 2) the built-in transformation from the 7^{th} to the 8^{th} (or output) layer in the CNN, to estimate the categorization probabilities from the decoded semantic representation. The first part of the model was trained with the fMRI data during the training movie; the second part was established by retraining the CNN for image classification into 15 categories. After training, we evaluated the decoding performance with the testing movie. Fig. 2.9 shows the top-5 decoded categories, ordered by their descending probabilities, in comparison with the true categories shown in red. On average, the top-1/top-2/top-3 accuracies were about 48%/65%/72%, significantly better than the chance levels (6.9%/14.4%/22.3%) (Table 2.1). These results confirm that cortical fMRI activity contained rich categorical representations, as previously shown elsewhere [60, 61, 87]. Along with visual reconstruction, direct categorization yielded textual descriptions. As an example, a flying bird seen by a subject was not only reconstructed as a bird-like image, but also described as a word “bird” (see the first frame in Figs. 2.8 & 2.9).

2.3.8 Cross-subject encoding and decoding

Different subjects’ cortical activity during the same training movie were generally similar, showing significant inter-subject reproducibility of the fMRI signal ($p<0.01$, t-test, Bonferroni correction) for 82% of the locations within visual cortex (Fig. 2.10.a). This lent support to the feasibility of neural encoding and decoding across different subjects – predicting and decoding one subject’s fMRI activity with the encoding/decoding models trained with data from another subject. Indeed, it was found that the encoding models could predict cortical fMRI responses

across subjects with still significant, yet reduced, prediction accuracies for most of the visual cortex (Fig. 2.10.b). For decoding, low-level feature representations (through the 1st layer in the CNN) could be estimated by inter-subject decoding, yielding reasonable accuracies only slightly lower than those obtained by training and testing the decoding models with data from the same subject (Fig. 2.10.c). The semantic categorization by inter-subject decoding yielded top-1 through top-3 accuracies as 24.9%, 40.0% and 51.8%, significantly higher than the chance levels (6.9%, 14.4% and 22.3%), although lower than those for intra-subject decoding (47.7%, 65.4%, 71.8%) (Fig. 2.10.d and Table 2.1). Together, these results provide evidence for the feasibility of establishing neural encoding and decoding models for a general population, while setting up the baseline for potentially examining the disrupted coding mechanism in pathological conditions.

2.4 Discussion

This study extends a growing body of literature in using deep learning models for understanding and modeling cortical representations of natural vision [8-10, 22, 30, 55, 56]. In particular, it generalizes the use of convolutional neural network to explain and decode widespread fMRI responses to naturalistic movie stimuli, extending the previous findings obtained with static picture stimuli. This finding lends support to the notion that cortical activity underlying dynamic natural vision is largely shaped by hierarchical feedforward processing driven towards object recognition, not only for the ventral stream, but also for the dorsal stream, albeit to a lesser degree. It sheds light on the object representations along the dorsal stream.

Despite its lack of recurrent or feedback connections, the CNN enables a fully computable predictive model of cortical representations of any natural visual input. The voxel-wise encoding model enables the visualization of single-voxel representation, to reveal the distinct functions of individual cortical locations during natural vision. It further creates a high-throughput computational workbench for synthesizing cortical responses to natural pictures, to enable cortical mapping of category representation and selectivity without running fMRI experiments. In addition, the CNN also enables direct decoding of cortical fMRI activity to estimate the feature representations in both visual and semantic spaces, for real-time visual reconstruction and semantic categorization of natural movie stimuli. In summary, the CNN-based encoding and decoding models, trained with hours of fMRI data during movie viewing, establish a computational account of feedforward cortical activity throughout the entire visual cortex and across all levels of

processing. Subsequently, we elaborate the implications from methodology, neuroscience, and artificial intelligence perspectives.

2.4.1 CNN predicts nonlinear cortical responses throughout the visual hierarchy

The brain segregates and integrates visual input through cascaded stages of processing. The relationship between the visual input and the neural response bears a variety of nonlinearity and complexity [2]. It is thus impossible to hand-craft a general class of models to describe the neural code for every location, especially for those involved in the mid-level processing. The CNN accounts for natural image statistics with a hierarchy of nonlinear feature models learned from millions of labeled images. The feature representations of any image or video can be automatically extracted by the CNN, progressively ranging from the visual to semantic space. Such feature models offer a more convenient and comprehensive set of predictors to explain the evoked fMRI responses, than are manually defined [44, 60]. For each voxel, the encoding model selects a subset from the feature bank to best match the voxel response with a linear projection. This affords the flexibility to optimally model the nonlinear stimulus-response relationship to maximize the response predictability for each voxel.

In this study, the model-predictable voxels cover nearly the entire visual cortex (Fig. 2.3.a), much beyond the early visual areas predictable with Gabor or motion filters[24, 26, 88], or with manually-defined categorical features [44, 60]. It is also broader than the incomplete ventral stream previously predicted by similar models trained with limited static pictures [8, 30, 56]. The difference is likely attributed to the larger sample size of our training data, conveniently afforded by video stimuli rather than picture stimuli. The PCA-based feature-dimension reduction also contributes to more robust and efficient model training. However, the encoding models only account for a fraction of the explainable variance (Fig. 2.4), and hardly explain the most lateral portion of early visual areas (Fig. 2.3.a). This area tends to have a lower SNR, showing lower intra-subject reproducibility (Fig. 2.2.a) or explainable variance (Fig. 2.4.a). The same issue also appears in other studies [8, 51], whereas the precise reason remains unclear.

Both the ventral stream and the CNN are presumably driven by the same goal of object recognition. Hence, it is not surprising that the CNN is able to explain a significant amount of cortical activity along the ventral stream, in line with prior studies [8-10, 30]. It further confirms the paramount role of feedforward processing in object recognition and categorization [89].

What is perhaps surprising is that the CNN also predicts dorsal-stream activity. The ventral-dorsal segregation is a classical principle of visual processing: the ventral stream is for perception ("what"), and the dorsal stream is for action ("where") [90]. As such, the CNN aligns with the former but not the latter. However, dorsal and ventral areas are inter-connected, allowing cross-talk between the pathways [91]. The dichotomy of visual streams is debatable [76]. Object representations exist in both ventral and dorsal streams with likely dissociable roles in visual perception [75]. Our study supports this notion. The hierarchical features extracted by the CNN are also mapped onto the dorsal stream, showing a representational gradient of complexity, as does the ventral stream. Nevertheless, the CNN accounts for a higher portion of the explainable variance for the ventral stream than for the dorsal stream (Fig. 2.4). We speculate that motion and attention sensitive areas in the dorsal stream require more than feedforward perceptual representations, while involving recurrent and feedback connections [92] that are absent in the CNN. In this regard, we would like to clarify that the CNN in the context of this paper is driven by image recognition and extracts spatial features, in contrast to 3-D convolutional network trained to extract spatiotemporal features for action recognition [93], which was another plausible model for the dorsal-stream activity [53].

2.4.2 Visualization of single-voxel representation reveals functional specialization

An important contribution of this study is the method for visualizing single-voxel representation. It reveals the specific pixel pattern from the visual input that gives rise to the response at the voxel of interest. The method is similar to those for visualizing the representations of individual units in the CNN [43, 68]. Extending from CNN units to brain voxels, it is helpful to view the encoding models as an extension of the CNN, where units are linearly projected onto voxels through voxel-wise encoding models. By this extension, the pixel pattern is optimized to maximize the model prediction of the voxel response, revealing the voxel's representation of the given visual input, using a combination of masking [64] and gradient [20, 66, 68] based methods. Here, visualization is tailored to each voxel, instead of each unit or layer in the CNN, setting it apart from prior studies [8, 20, 43, 68].

Utilizing this visualization method, one may reveal the distinct representations of the same visual input at different cortical locations. As exemplified in Fig. 2.6, visualization uncovers the increasingly complex and category-selective representations for locations running downstream

along the visual pathways. It offers intuitive insights into the distinct functions of different locations, e.g. the complementary representations at FFA and PPA. Although we focus on the methodology, our initial results merit future studies for more systematic characterization of the representational differences among voxels in various spatial scales. The visualization method is also applicable to single or multi-unit activity, to help understand the localized responses of neurons or neuronal ensembles [9].

2.4.3 High-throughput computational workbench for studying natural vision

The CNN-based encoding models, trained with a large and diverse set of natural movie stimuli, can be generalized to other novel visual stimuli. Given this generalizability, one may use the trained encoding models to predict and analyze cortical responses to a large number of natural pictures or videos, much beyond what is practically doable with fMRI scans. As such, the encoding models constitute a high-throughput computational workbench for studying the neural representations of natural vision. As shown here and elsewhere [30], this workbench is immediately usable for mapping categorical representation, contrast, and selectivity, to yield novel hypotheses for further experimental investigations. Open-access software platform is much desirable to further leverage this potential.

2.4.4 Direct visual reconstruction of a natural movie

For decoding cortical activity, the CNN enables direct reconstruction of natural movies. It does not require any comparison between the observed activity pattern and those generated by or predicted from candidate pictures. This sets our method apart from multivariate pattern analysis [17, 18, 23] and encoding-model-based decoding [24-26]. In particular, Nishimoto et al. (2011) published the first, and to date the only, attempt to reconstruct natural movies. They used a "try-and-error" strategy: searching a huge prior set of videos for the most likely stimuli that would match the measured cortical activity through model prediction by the encoding models. Arguably, this strategy is difficult to scale up because it is impossible for any prior set to be fully inclusive. The identification or reconstruction accuracy is dependent on and biased by the samples in the prior set. The need for a large prior set is also computationally expensive, limiting the decoding efficiency.

A prior study [85] tried to avoid these limitations. In that study, the fMRI signals were used to estimate the contrast of local image bases, which in turn were combined to directly reconstruct small, simple, and binary images. While the method is not constrained or biased by any image prior, binary image bases are not suitable for describing natural image statistics even in the lowest level. Also note that the decoding models in that study were trained with a small set of random images, and tested with simple letters and shapes. However, realistic visual input is complex and dynamic, and natural vision involves salience and attention [83, 84]. Such complexity is unlikely captured by random and binary pixel patterns [52]. The overall strategy, as described in [85], is not readily usable to decode dynamic natural visual experiences.

Our decoding method does not require any prior set of candidate images, setting itself apart from the encoding-model-based decoding [26]. It also uses features learned from natural images, different from the method in [85]. The latter is important because the features in the CNN are biologically relevant [2] and capture information useful for perception [4]. In particular, the 1st layer includes features of orientation, contrast, edge, and color, forming a more informative basis set than binary image bases [85].

In this study, visual reconstruction was only based on the fMRI-decoded 1st-layer features. Although the feature representations from other layers could also be estimated with comparable accuracies [7], combining the estimated features from all layers did not improve visual reconstruction. Multiple reasons are conceivable. Higher layers contain more abstract information and contribute less to the specific pixel patterning [94]. The De-CNN reverses the CNN with approximation, especially at the un-pooling step. As a result, the decoding errors cascade down the CNN, causing accumulated errors in the reconstructed pixels.

In this study, the fMRI-decoded visual reconstruction emphasized foreground and suppressed background (Fig. 2.8). This intriguing finding is likely attributable to the effects from both bottom-up salience [83] and top-down attention [84]. The CNN captures visual salience [20, 95], but has no mechanism for top-down attention. It thus helps to dissociate the salience vs. attention effects. To explore the effects from salience but not attention, we applied the decoding model to the fMRI signals predicted by the voxel-wise encoding models. As in Supplementary Fig. 2.9 in [7], the resulting visual reconstruction also highlighted the foreground objects. It suggests that visual salience is captured by the CNN and indeed contributes to the foreground selectivity. However, decoding of the measured fMRI signals revealed even more focal emphases on

foreground objects. Therefore, in addition to bottom-up salience, there are other selection mechanisms, likely top-down attention that shape the fMRI responses during movie viewing.

2.4.5 Direct decoding of semantic representations and categorization

This study also demonstrates the value of using the CNN to directly decode and categorize semantic representations. The CNN contains a semantic space in its 2nd highest layer. It supports object recognition in the output layer with either finely or coarsely defined categories, and is even transferrable to other vision tasks [50]. Hence, it represents a generalizable semantic space, emerging progressively from the visual features in the lower levels. The decoding model allows us to directly estimate the representation in this semantic space for arbitrary natural stimuli. The decoded semantic representation is generalizable and transferable, and independent of the definition of categories, unlike the categorical decoding method recently reported elsewhere [87].

In addition, the semantic space in the CNN can be readily converted to human-defined categorical labels, by training a classifier to match the semantic representation to the label. It effectively translates a vector representation to a word, and allows the textual interpretation of brain activity. The classifier can be trained without redefining the semantic space, by only retraining the CNN's output layer with labeled images. So, the classifier is separate from the decoding model. This offers interesting extensions of the current decoding capabilities. One may utilize the ever-expanding labeled images to set up various interpretations of the semantic representations decoded from brain activity.

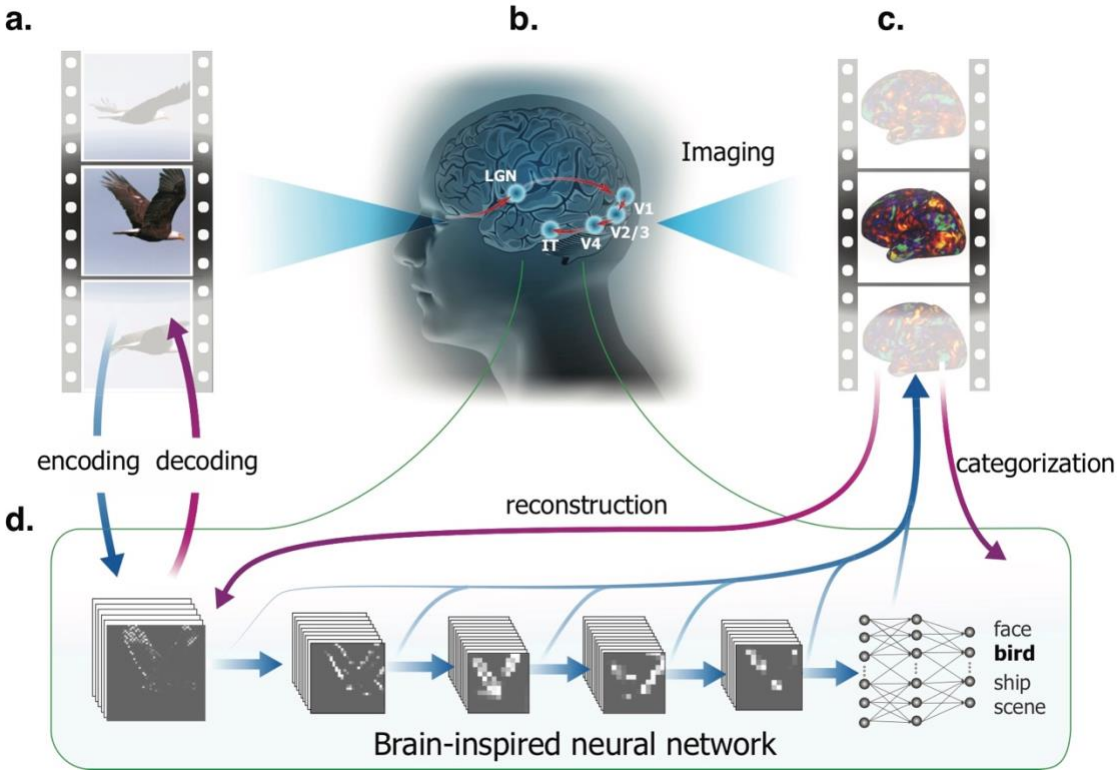

Figure 2.1 **Neural encoding and decoding through a deep-learning model.** When a person is seeing a film (a), information is processed through a cascade of cortical areas (b), generating fMRI activity patterns (c). A deep CNN is used here to model cortical visual processing (d). This model transforms every movie frame into multiple layers of features, ranging from orientations and colors in the visual space (the first layer) to object categories in the semantic space (the eighth layer). For encoding, this network serves to model the nonlinear relationship between the movie stimuli and the response at each cortical location. For decoding, cortical responses are combined across locations to estimate the feature outputs from the first and seventh layer. The former is deconvolved to reconstruct every movie frame, and the latter is classified into semantic categories.

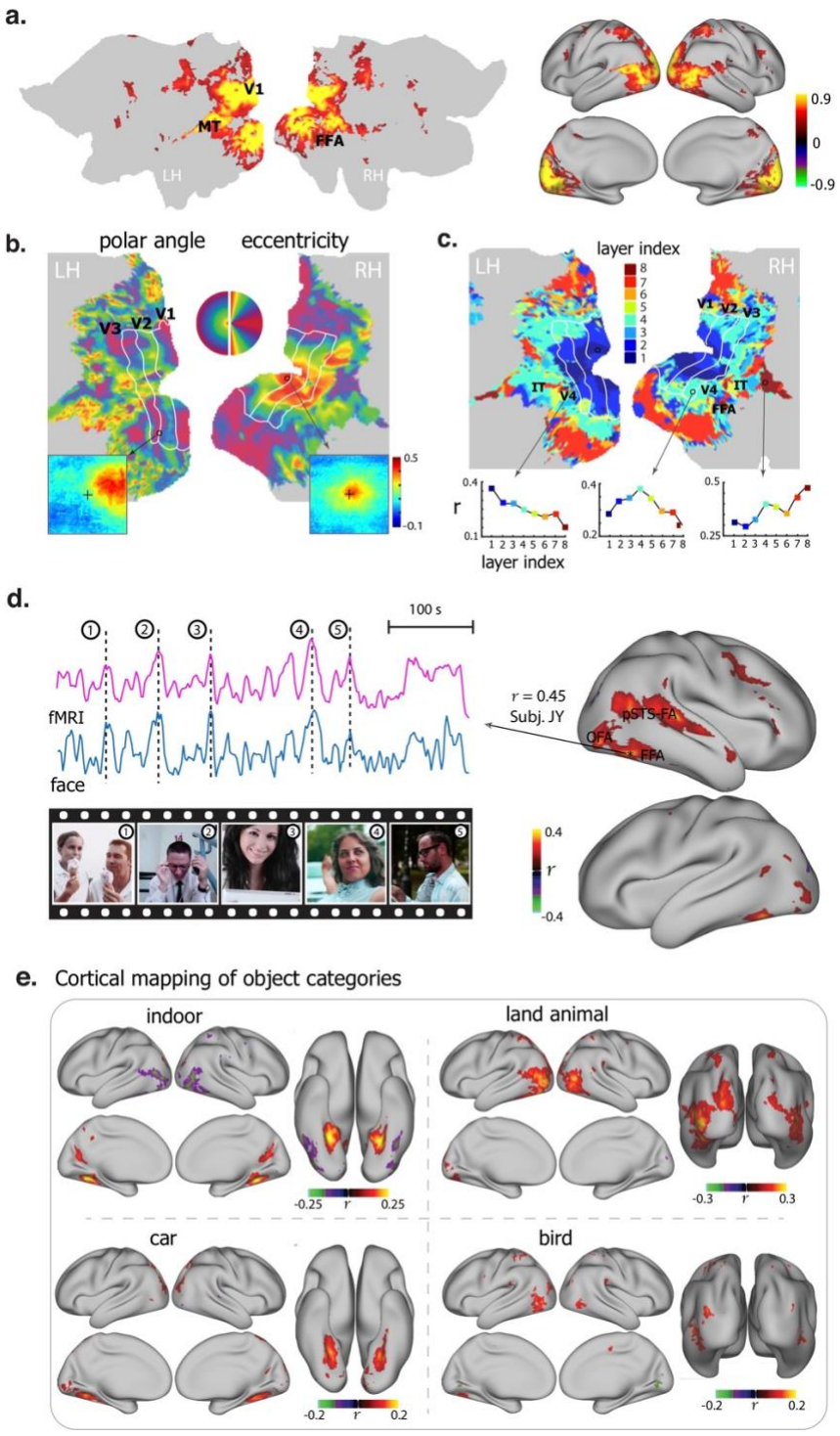

Figure 2.2 **Functional alignment between the visual cortex and the CNN during natural vision.** (a) Cortical activation. The maps show the cross correlations between the fMRI signals obtained during 2 repetitions of the identical movie stimuli. (b) "Retinotopic mapping". Cortical representations of the polar angle (left) and eccentricity (right), quantified for the receptive-field center of every cortical location, are shown on the flattened cortical surfaces. The bottom insets show the receptive fields of 2 example locations from V1 (right) and V3 (left). The V1/V2/V3

borders defined from conventional retinotopic mapping are overlaid for comparison. (c) “Hierarchical mapping”. The map shows the index to the CNN layer most correlated with every cortical location. For 3 example locations, their correlations with different CNN layers are displayed in the bottom plots. (d) “Co-activation of FFA in the brain and the ‘Face’ unit in the CNN”. The maps on the right show the correlations between cortical activity and the output time series of the “Face” unit in the eighth layer of CNN. On the left, the fMRI signal at a single voxel within the FFA is shown in comparison with the activation time series of the “Face” unit. Movie frames are displayed at 5 peaks co-occurring in both time series for 1 segment of the training movie. The selected voxel was chosen since it had the highest correlation with the “face” unit for other segments of the training movie, different from the one shown in this panel. (e) “Cortical mapping of other 4 categories”. The maps show the correlation between the cortical activity and the outputs of the eighth-layer units labeled as “indoor objects”, “land animals”, “car”, “bird”. See Supplementary Figs 2, 3, and 4 in [7] for related results from individual subjects.

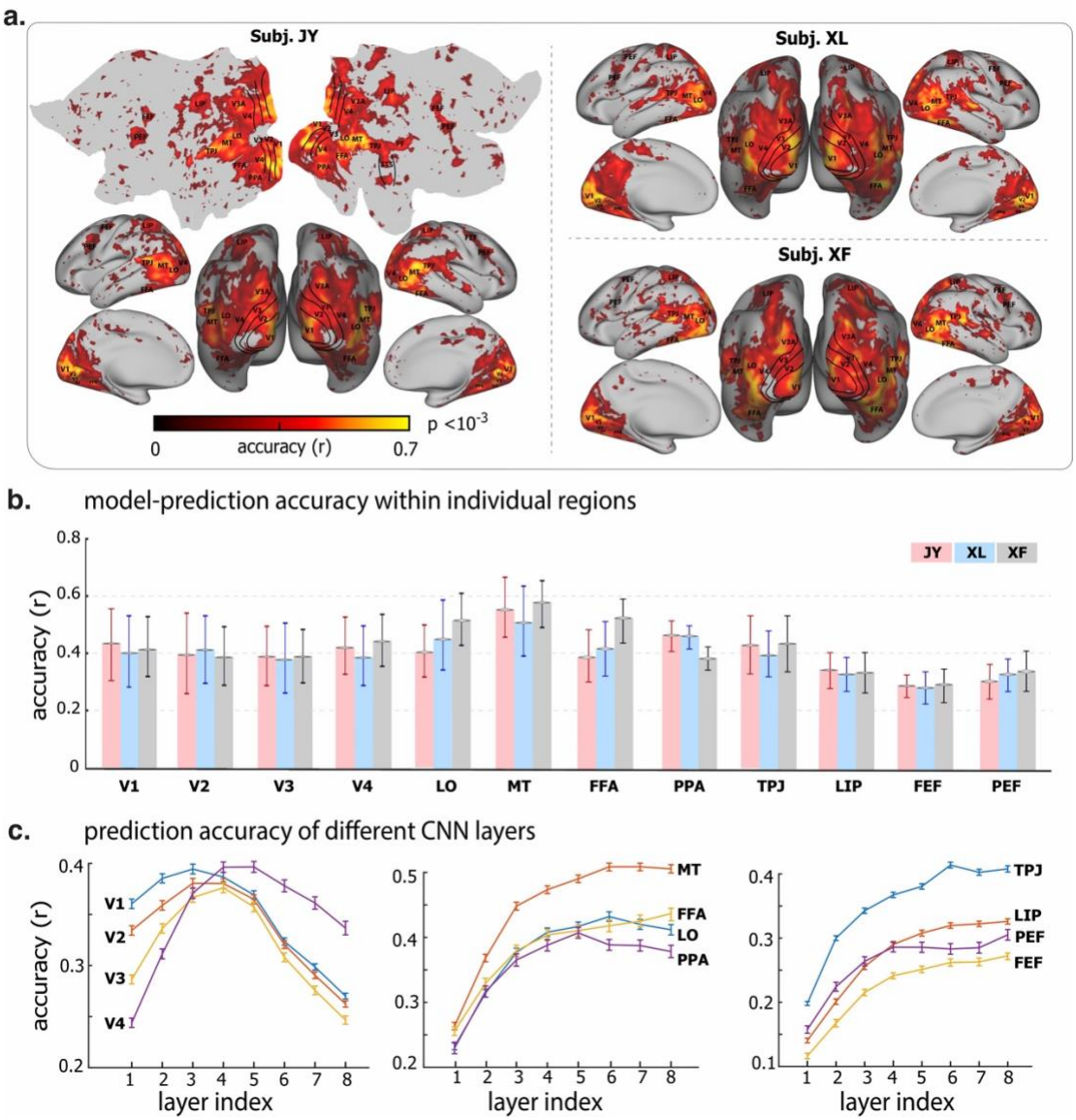

Figure 2.3 **Cortical predictability given voxel-wise encoding models.** (a) Accuracy of voxel-wise encoding models in predicting the cortical responses to novel natural movie stimuli, which is quantified as the Pearson correlation between the measured and the model-predicted responses during the testing movie. (b) Prediction accuracy within regions of interest (ROIs) for 3 subjects. For each ROI, the prediction accuracy is summarized as the mean ± std correlation for all voxels within the ROI. (c) Prediction accuracy for different ROIs by different CNN layers. For each ROI, the prediction accuracy was averaged across voxels within the ROI, and across subjects. The curves represent the mean, and the error bars stand for the standard error.

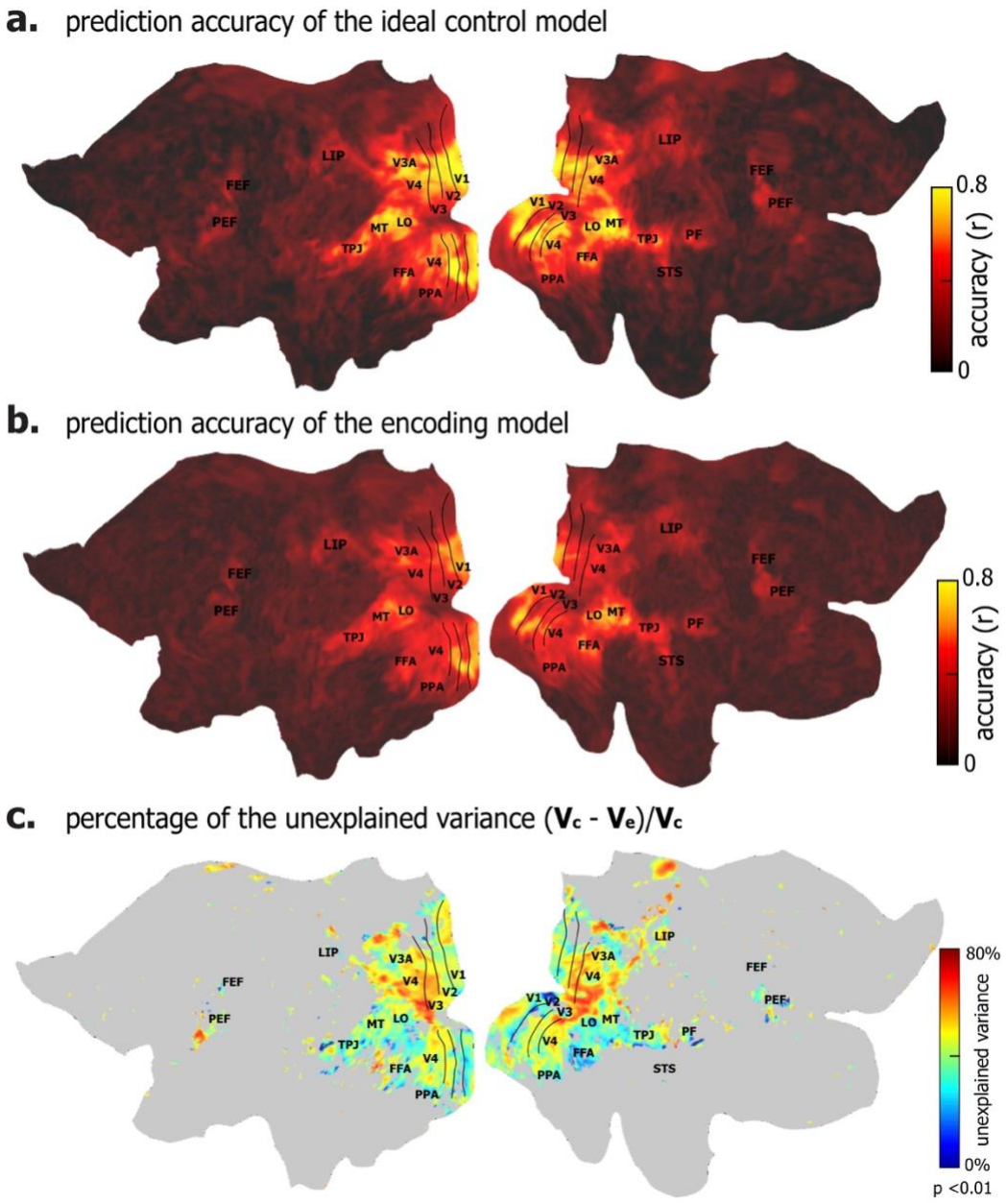

Figure 2.4 **Explained variance of the encoding models.** (a) Prediction accuracy of the ideal control model (average across subjects). It defines the potentially explainable variance in the fMRI signal. (b) Prediction accuracy of the CNN-based encoding models (average across subjects). (c) The percentage of the explainable variance that is not explained by the encoding model. Vc denotes the potentially explainable variance and Ve denotes the variance explained by the encoding model. Note that this result was based on movie-evoked responses averaged over 5 repetitions of the testing movie, while the other 5 repetitions were used to define the ideally explainable variance. This was thus distinct from other figures, which were based on the responses averaged over all 10 repetitions of the testing movie.

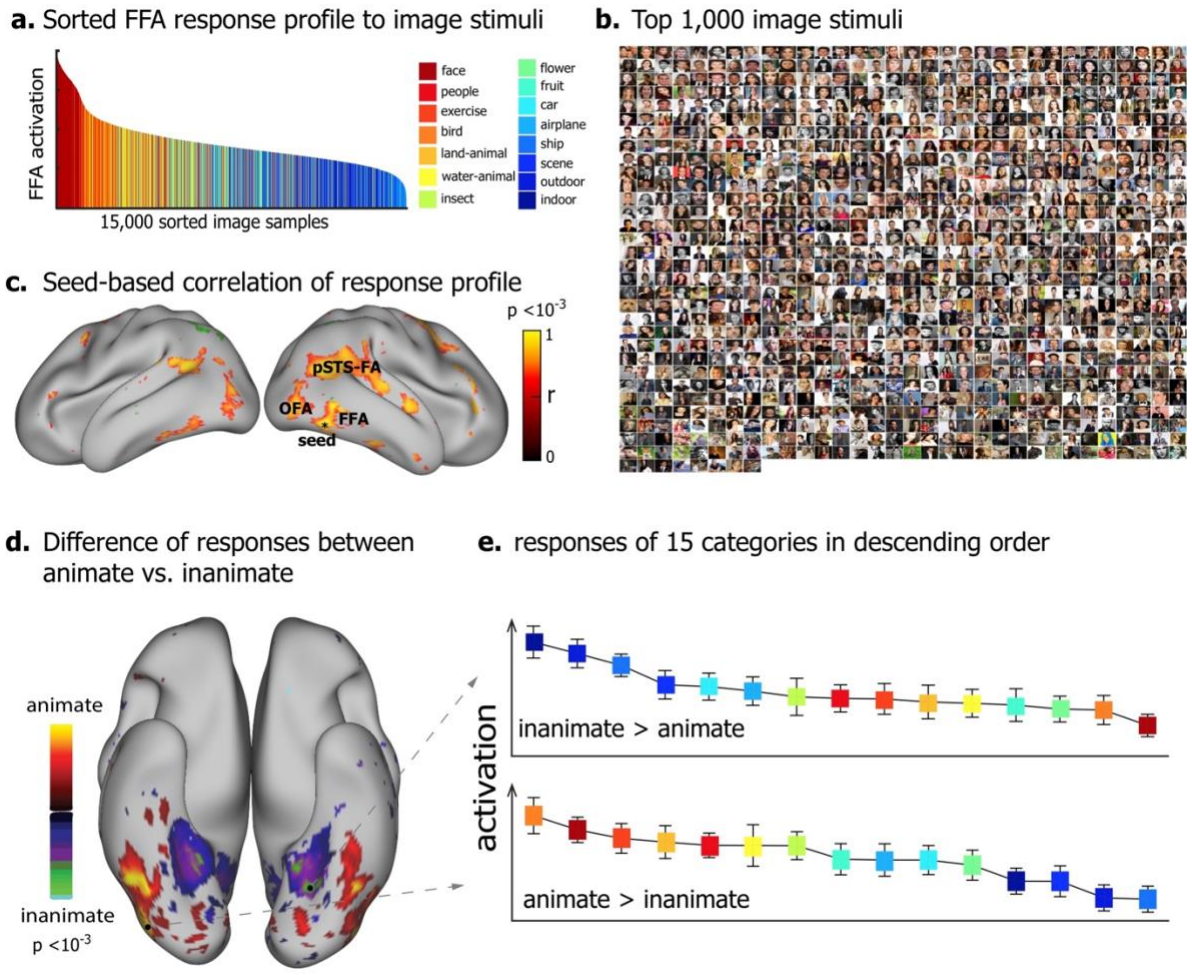

Figure 2.5 **Cortical representations of single-pictures or categories.** (a) The model-predicted response profile at a selected voxel in FFA given 15 000 natural pictures from 15 categories, where the selected voxel had the highest prediction accuracy when the encoding model was evaluated using the testing movie. The voxel's responses are sorted in descending order. (b) The top-1 000 pictures that generate the greatest responses at this FFA voxel. (c) Correlation of the response profile at this "seed" voxel with those at other voxels ($P < 0.001$, Bonferroni correction). (d) The contrast between animate versus inanimate pictures in the model-predicted responses (2-sample t-test, $P < 0.001$, Bonferroni correction). (e) The categorical responses at 2 example voxels. These 2 voxels show the highest animate and inanimate responses, respectively. The colors correspond to the categories in (a). The results are from Subject JY, see Supplementary Fig. 5 in [7] for related results from other subjects.

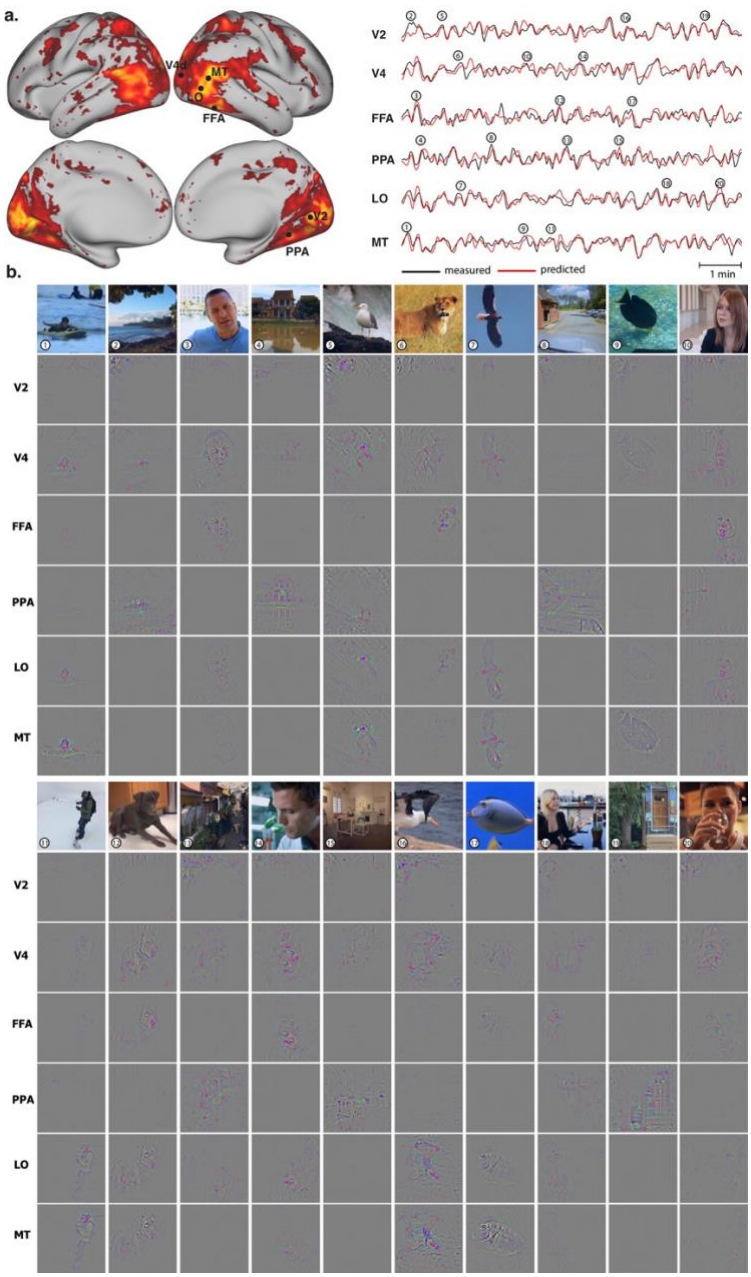

Figure 2.6 **Neural encoding models predict cortical responses and visualize functional representations at individual cortical locations.** (a) Cortical predictability for subject JY, same as Fig. 2.3a. The measured (black) and predicted (red) response time series are also shown in comparison for 6 locations at V2, V4, LO, MT, PPA, and FFA. For each area, the selected location was the voxel within the area where the encoding models yielded the highest prediction accuracy during the testing movie (b) Visualizations of the 20 peak responses at each of the 6 locations shown in (a). The presented movie frames are shown in the top row, and the corresponding visualizations at 6 locations are shown in the following rows. The results are from Subject JY, see Supplementary Figs 6 and 7 in [7] for related results from other subjects.

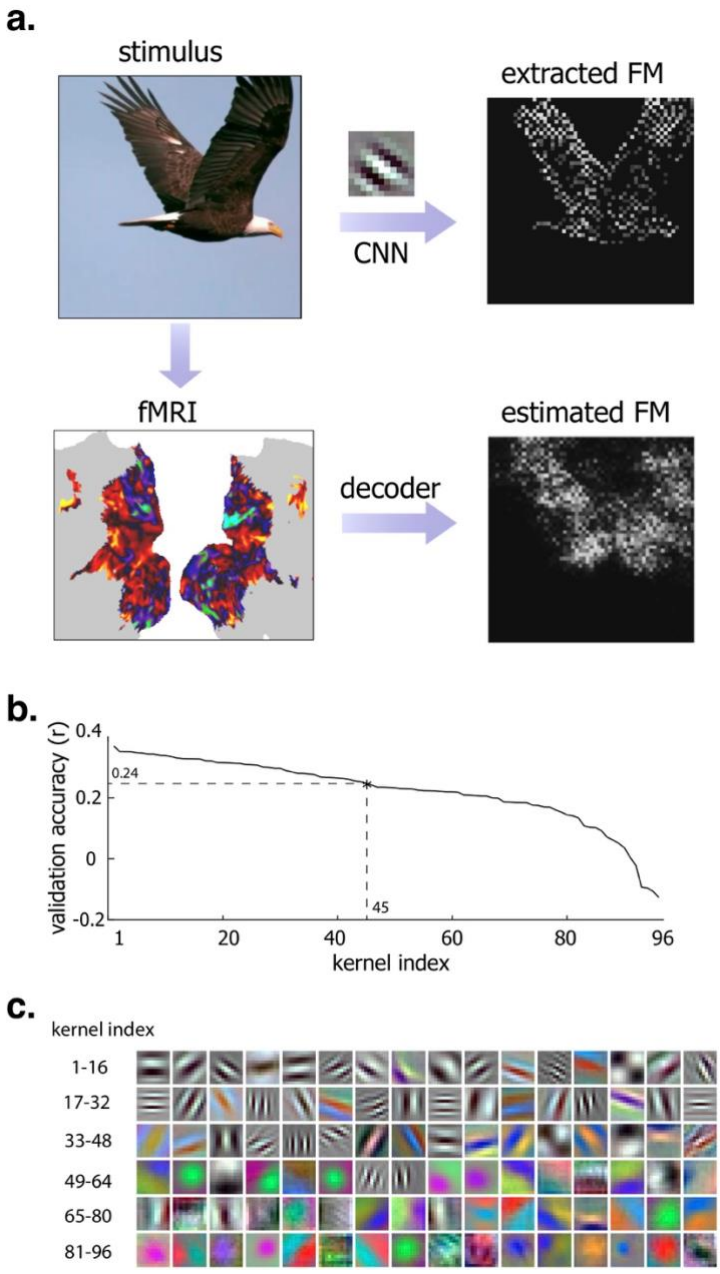

Figure 2.7 **fMRI-based estimation of the first-layer feature maps (FM).** (a) For each movie frame, the feature maps extracted from the kernels in the first CNN layer were estimated from cortical fMRI data through decoders trained with the training movie. For an example movie frame (flying eagle) in the testing movie, its feature map extracted with an orientation-coded kernel revealed the image edges. In comparison, the feature map estimated from fMRI was similar, but blurrier. (b) The estimation accuracy for all 96 kernels, given cross-validation within the training data. The accuracies were ranked and plotted from the highest to lowest. Those kernels with high accuracies ($r > 0.24$) were selected and used for reconstructing novel natural movies in the testing phase. (c) 96 kernels in the first layer are ordered in a descending manner according to their cross-validation accuracy.

Figure 2.8 **Reconstruction of a dynamic visual experience.** For each row, the top shows the example movie frames seen by 1 subject; the bottom shows the reconstruction of those frames based on the subject's cortical fMRI responses to the movie. See Movie 1 in [7] for the reconstructed movie.

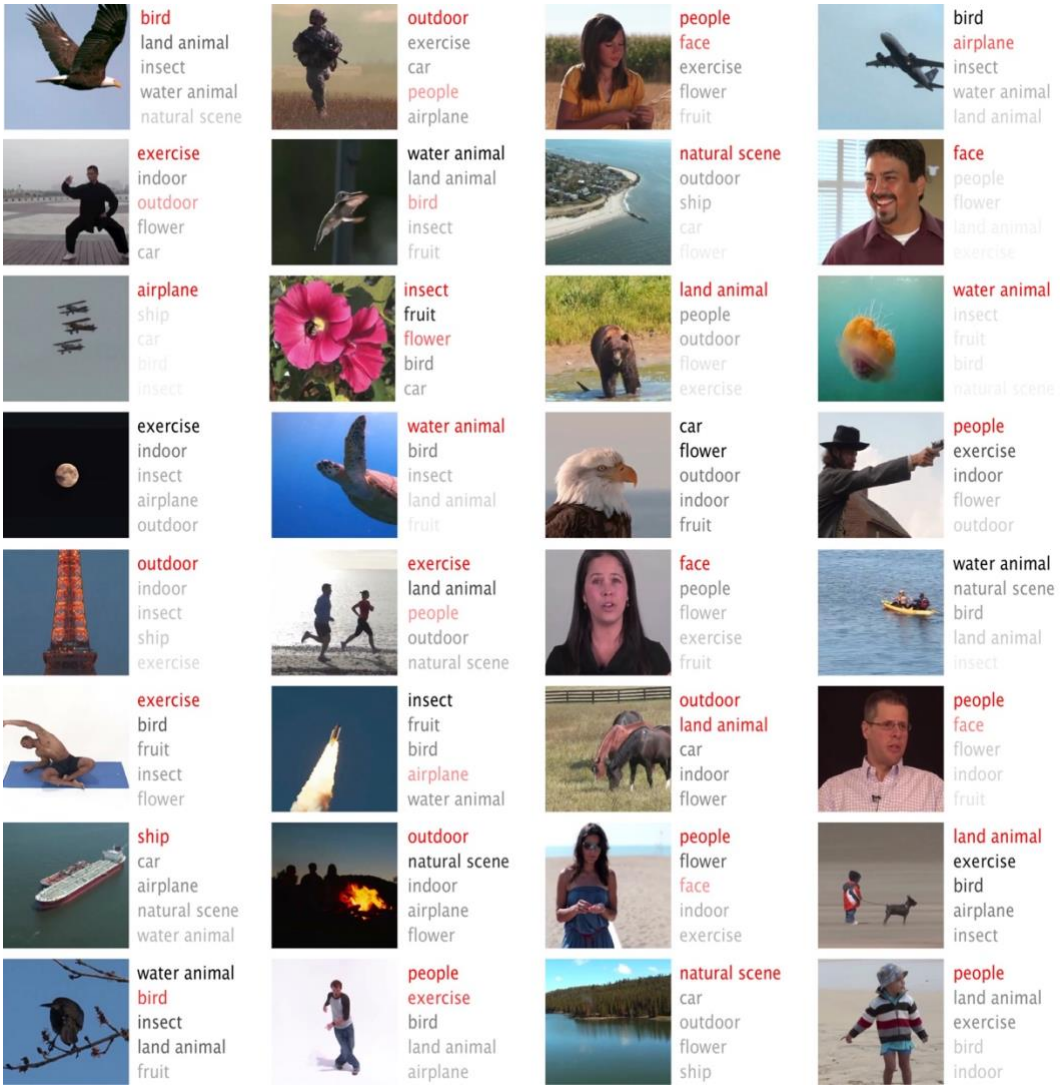

Figure 2.9 **Semantic categorization of natural movie stimuli.** For each movie frame, the top-5 categories determined from cortical fMRI activity are shown in the order of descending probabilities from the top to the bottom. The probability is also color coded in the gray scale with the darker gray indicative of higher probability. For comparison, the true category labeled by a human observer is shown in red. Here, we present the middle frame of every continuous video clip in the testing movie that could be labeled as one of the pre-defined categories. See Movie 1 in [7] for all other frames.

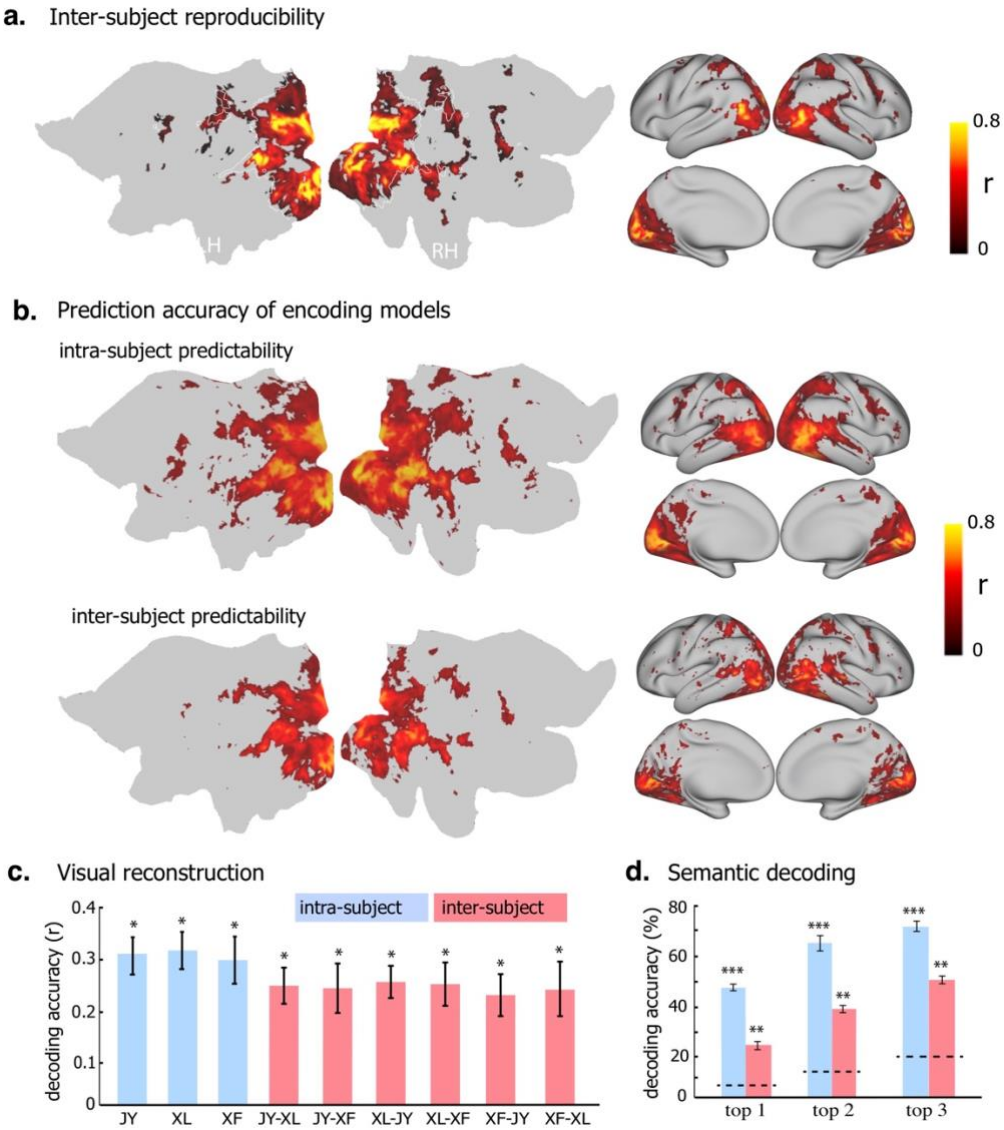

Figure 2.10 **Encoding and decoding within vs. across subjects.** (a) Average inter-subject reproducibility of fMRI activity during natural stimuli. (b) Cortical response predictability with the encoding models trained and tested for the same subject (i.e., intra-subject encoding) or for different subjects (i.e., inter-subject encoding). (c) Accuracy of visual reconstruction by intra-subject (blue) vs. inter-subject (red) decoding for 1 testing movie. The y-axis indicates the spatial cross correlation between the fMRI- estimated and CNN-extracted feature maps for the first layer in the CNN. The x-axis shows multiple pairs of subjects (JY, XL, and XF). The first subject indicates the subject from whom the decoder was trained; the second subject indicates the subject for whom the decoder was tested. (d) Accuracy of categorization by intra-subject (blue) vs. inter-subject (red) decoding. The top-1, top-2 and top-3 accuracy indicates the percentage by which the true category is within the first, second, and third most probable categories predicted from fMRI, respectively. For both (c) and (d), the bar height indicates the average prediction accuracy; the error bar indicates the standard error of the mean; the dashed lines are chance levels. (*P < 10−4, **P < 10−10, ***P < 10−50). See Movie 2 for the reconstructed movie on the basis of inter-subject decoding.

Table 2.1 Three sub-tables show the top-1, top-2 and top-3 accuracies of categorizing individual movie frames by using decoders trained with data from the same (intra-subject) or different (inter-subject) subject. Each row shows the categorization accuracy with the decoder trained with a specific subject's training data; each column shows the categorization accuracy with a specific subject's testing data and different subjects' decoders. The accuracy was quantified as the percentage by which individual movie frames were successfully categorized as one of the top-1, top-2, or top-3 categories. The accuracy was also quantified as a fraction number (shown next to the percentage number): the number of correctly categorized frames over the total number of frames that could be labeled by the 15 categories (N=214 for one 8-min testing movie).

Decoding accuracy for the semantic descriptions of a novel movie

	train \ test	subject 1	subject 2	subject 3
top 1	subject 1	42.52% (91/214)	24.30% (52/214)	23.83% (51/214)
	subject 2	20.09% (43/214)	50.47% (108/214)	22.90% (49/214)
	subject 3	24.77% (53/214)	33.64% (72/214)	50.00% (107/214)
top 2	subject 1	59.81% (128/214)	41.12% (88/214)	43.93% (94/214)
	subject 2	35.51% (76/214)	70.09% (150/214)	35.98% (77/214)
	subject 3	41.12% (88/214)	42.06% (90/214)	66.36% (142/214)
top 3	subject 1	67.76% (145/214)	55.14% (118/214)	53.27% (114/214)
	subject 2	48.13% (103/214)	74.77% (160/214)	50.93% (109/214)
	subject 3	50.93% (109/214)	52.34% (112/214)	72.90% (156/214)

3. DEEP NEURAL NETWORK PREDICTS CORTICAL REPRESENTATION AND ORGANIZATION OF VISUAL FEATURES FOR RAPID CATEGORIZATION

*Modified and formatted for dissertation from the article published in *Scientific Report* [27]

3.1 Introduction

The visual cortex is capable of rapid categorization of visual objects[96, 97]. This ability is attributable to cortical representation and organization of object information[2, 98]. In the ventral temporal cortex, object representations are topologically organized[99], spanning a high-dimensional space [100] and being largely invariant against low-level appearance[96, 101]. Knowledge about object categories is also represented in dorsal visual areas[75, 102, 103] or even beyond the visual cortex[104] where non-visual attributes of objects are coded[105, 106]. In addition to their distributed representations[80, 107], object attributes are hierarchically organized and progressively emerge from visual input[2]. It is thus hypothesized that the brain categorizes visual objects based on their attributes represented in multiple stages of visual processing[99, 106].

To understand the basis of object categorization, it is desirable to map cortical representations of as many objects from as many categories as possible. The resulting maps provide the stimulus-response samples to address the representational structure that enables the brain to categorize or differentiate visual objects. Many studies have used functional magnetic resonance imaging (fMRI) to map brain activations with category-specific images[57, 77, 80, 105, 108, 109]. Although such approaches are valuable for studying object categorization, it is expensive to cover many objects or categories in experiments, and it is arguably difficult to extrapolate experimental findings to new objects or categories. Moreover, object representations in the voxel space do not directly reveal the neural computation that give rise to such representations. It is also desirable to develop a model of hierarchical visual processing[110] to be able to explain (or predict) cortical representations of visual objects with (or without) experimental data.

Advances in deep learning[4] have established a range of deep neural networks (DNN) inspired by the brain itself[2, 3]. Such models have been shown to be able to achieve human-level performance in object classification, segmentation, and tracking[4]. On the basis of DNNs,

encoding models could be built to predict cortical responses to natural images[8-10, 22, 30] or videos[7, 53]. As the accuracies of predicted responses were high and robust in the entire visual cortex[7], DNN-based encoding models are arguably advantageous than other models that only account for either the lowest[24, 26] or highest[60] level of visual processing.

Recent studies also suggest that DNN-based encoding models may be generalized to new images or videos[7-9, 30]. In this sense, the models provide a platform to simulate cortical representations of in principle infinite exemplars of a large number of object categories[7, 30], beyond what is experimentally attainable[77, 79, 111-113]. In addition, DNN views an image as a set of hierarchically organized features, rather than as a pixel array. The features are learned from millions of images to model image statistics in different levels of abstraction[4]. The learned features are much richer and more fine-grained than what may be intuitively defined (by humans) as the mid-level features. Through DNN-based encoding models, it is plausible to map object representations of specific features from each layer in DNN, allowing object categorization to be addressed at each level of visual processing.

Extending from recent studies[7-10, 22, 30], we used a deep residual network (ResNet)[21] to define, train, and test a generalizable, predictive, and hierarchical model of natural vision by using extensive fMRI data from humans watching >10 hours of natural videos. Taking this predictive model as a "virtual" fMRI scanner, we synthesized the cortical response patterns with 64,000 natural pictures including objects from 80 categories, and mapped cortical representations of these categories with high-throughput. We evaluated the category selectivity at every voxel in the visual cortex, compared cortical representational similarity with their semantic relationships, and evaluated the contributions from different levels of visual features to the cortical organization of categories. Consistent but complementary to prior experimental studies[57, 60, 80, 105, 114-119], this study used a model-based computational strategy to study how cortical representations of various levels of object knowledge sub-serve categorization.

3.2 Methods and Materials

3.2.1 Experimental data

We used and extended the human experimental data from our previous study[7], according to experimental protocols approved by the Institutional Review Board at Purdue University with

informed consent from all human subjects prior to their participation. All methods were performed in accordance with the relevant guidelines and regulations. Briefly, the data included the fMRI scans from three healthy subjects (Subject 1, 2, 3, all female) when watching natural videos. For each subject, the video-fMRI data were split into two independent datasets: one for training the encoding model and the other for testing it. For Subject 2 & 3, the training movie included 2.4 hours of videos; the testing movie included 40 minutes of videos; the training movie was repeated twice, and the testing movie was repeated ten times. For Subject 1, the training movie included not only those videos presented to Subject 2 and 3, but also 10.4 hours of new videos. The new training movie was presented only once. The movie stimuli included a total of ~9,300 video clips manually selected from *YouTube*, covering a variety of real-life visual experiences. All video clips were concatenated in a random sequence and separated into 8-min sessions. Every subject watched each session of videos (field of view: $20.3^{o}\times20.3^{o}$) through a binocular goggle with the eyes fixating at a central cross ($0.8^{o}\times0.8^{o}$). During each session, whole-brain fMRI scans were acquired with 3.5 mm isotropic resolution and 2 s repetition time in a 3-T MRI system by using a single-shot, gradient-recalled echo-planar imaging sequence (38 interleaved axial slices with 3.5 mm thickness and 3.5 × 3.5 mm2 in-plane resolution, TR = 2000 ms, TE = 35 ms, flip angle = 78°, field of view = 22 × 22 cm2). Structural MRI data with T_1 and T_2 weighted contrast were also acquired with 1 mm isotropic resolution for every subject. The volumetric fMRI data were preprocessed and co-registered onto a standard cortical surface template[49]. For each cortical location, the 4^{th}-order polynomial trend was removed from the fMRI signal. For training and testing encoding models (as described latter), the fMRI signals were averaged over repetitions if there were multiple repeats and then standardized (i.e. remove the mean and normalize the variance). More details about the movie stimuli, data preprocessing and acquisition are described elsewhere [7].

3.2.2 Deep residual network

In line with previous studies[7-10, 22, 30, 56], a feedforward deep neural network (DNN) was used to model the cortical representations of natural visual stimuli. Here, we used a specific version of the DNN known as the deep residual network (ResNet), which had been pre-trained to categorize natural pictures with the state-of-the-art performance[21]. In the ResNet, 50 hidden layers of neuron-like computational units were stacked into a bottom-up hierarchy. The first layer encoded location and orientation-selective visual features, whereas the last layer encoded semantic

features that supported categorization. The layers in between encoded increasingly complex features through 16 residual blocks; each block included three successive layers and a shortcut directly connecting the input of the block to the output of the block[21]. Compared to the DNNs in prior studies[7-9, 22, 56, 120], the ResNet was much deeper and defined more fine-grained hierarchical visual features. The ResNet could be used to extract feature representations from any input image or video frame by frame. Passing an image into the ResNet yielded an activation value at each unit. Passing a video yielded an activation time series at each unit as the fluctuating representation of a given visual feature in the video.

3.2.3 Encoding models

For each subject, we trained an encoding model to predict each voxel's fMRI response to any natural visual stimuli[12], using a similar strategy as previously explored[7, 8, 30]. The voxel-wise encoding model included two parts: the first part was nonlinear, converting the visual input from pixel arrays into representations of hierarchical features through the ResNet; the second part was linear, projecting them onto each voxel's fMRI response. The encoding model used the features from 18 hidden layers in the ResNet, including the first layer, the last layer, and the output layer for each of the 16 residual blocks. For video stimuli, the time series extracted by each unit was convolved with a canonical hemodynamic response function (HRF) with the peak response at 4s, and down-sampled to match the sampling rate of fMRI, and then standardized (i.e. remove the mean and normalize the variance).

The feature dimension was reduced by applying principle component analysis (PCA) first to each layer and then to all layers in ResNet. The principal components of each layer were a set of orthogonal vectors that explained >99% variance of the layer's feature representations given the training movie. The layer-wise dimension reduction was expressed as equation (1).

$$\boldsymbol{f}_l(\mathbf{x}) = \boldsymbol{f}_l^o(\mathbf{x})\mathbf{B}_l \qquad (1)$$

where $\boldsymbol{f}_l^o(\mathbf{x})$ $(1 \times p_l)$ is the original feature representation from layer l given a visual input $\mathbf{x}$, $\mathbf{B}_l$ $(p_l \times q_l)$ consists of unitary columnar vectors that represented the principal components for layer l, $\boldsymbol{f}_l(\mathbf{x})$ $(1 \times q_l)$ is the feature representation after reducing the dimension from p_l to q_l.

Following the layer-wise dimension reduction, the feature representations from all layers were further reduced by using PCA to retain >99% variance across layers. The final dimension reduction was implemented as equation (2).

$$\boldsymbol{f}(\mathbf{x}) = \boldsymbol{f}_{1:L}(\mathbf{x})\mathbf{B}_{1:L} \qquad (2)$$

where $\boldsymbol{f}_{1:L}(\mathbf{x}) = \left[\frac{\boldsymbol{f}_1(\mathbf{x})}{\sqrt{p_1}}, \ldots, \frac{\boldsymbol{f}_L(\mathbf{x})}{\sqrt{p_L}}\right]$ is the feature representation concatenated across L layers, $\mathbf{B}_{1:L}$ consists of unitary principal components of the layer-concatenated feature representations of the training movie, and $\boldsymbol{f}(\mathbf{x})$ $(1 \times k)$ is the final dimension-reduced feature representation.

For the second part of the encoding model, a linear regression model was used to predict the fMRI response $r_v(\mathbf{x})$ at voxel v evoked by the stimulus $\mathbf{x}$ based on the dimension-reduced feature representation $\boldsymbol{f}(\mathbf{x})$ of the stimulus, as expressed by equation (3).

$$r_v(\mathbf{x}) = \boldsymbol{f}(\mathbf{x})\, \mathbf{w}_v + \varepsilon_v \qquad (3)$$

where $\mathbf{w}_v$ is a columnar vector of regression coefficients specific to voxel v, and ε_v is the error term. As shown in equation (4), L_2-regularized least-squares estimation was used to estimate $\mathbf{w}_v$ given the data during the training movie (individual frames were indexed by $i = 1, \cdots, N$), where the regularization parameter was determined based on nine-fold cross-validation.

$$\widehat{\mathbf{w}}_v = \underset{\mathbf{w}_v}{\arg\min}\, \frac{1}{N}\sum_{i=1}^{N}(r_v(\mathbf{x}_i) - \boldsymbol{f}(\mathbf{x}_i)\, \mathbf{w}_v)^2 + \lambda \| \mathbf{w}_v \|_2^2 \qquad (4)$$

After the above training, the voxel-wise encoding models were evaluated for their ability to predict the cortical responses to the novel testing movie (not used for training). The prediction accuracy was quantified as the temporal correlation (r) between the predicted and observed fMRI responses at each voxel given the testing movie. Since the testing movie included five distinct sessions, the prediction accuracy was evaluated separately for each session, and then averaged across sessions. The significance of the voxel-wise prediction accuracy was evaluated with a block-permutation test[59] (corrected at false discovery rate (FDR) $q < 0.01$), as used in our prior study [7].

We also evaluated the correspondence between the hierarchical layers in ResNet and the hierarchical cortical areas underlying different stages of visual processing, in line with previous studies[7, 9, 10, 22, 30, 53, 56]. For this purpose, we calculated the variance of the response at a voxel explained by the visual features in single layers. Specifically, the features extracted from the testing movie were kept only for one layer in the ResNet, while setting to zeros for all other layers. Through the voxel-wise encoding model, the variance (measured by R-squared) of the response explained by the single layer was calculated. For each voxel, we identified the best corresponding layer with the maximum explained variance and assigned its layer index to this voxel. The assigned layer index indicated the processing stage this voxel belonged to.

We also tested whether the deeper ResNet outperformed the shallower AlexNet[19] in predicting cortical responses to natural movies, taking the latter as the benchmark given its state-of-the-art encoding performance in prior studies[7, 8, 22]. For this purpose, we trained and tested similar encoding models based on the AlexNet with the same analysis of the same dataset. We compared the prediction accuracy between ResNet and AlexNet for regions of interest (ROIs) defined in an existing cortical parcellation[62], and further evaluated the statistical significance of their difference using a paired t-test ($p<0.001$) across all voxels within each ROI. Considering the noise in the data, we also calculated the noise ceiling of the predictability at each voxel. The noise ceiling indicated the maximum accuracy that a model could be expected to achieve given the level of noise in the testing data[121]. The noise and signal in fMRI were assumed to follow Gaussian distribution and the mean of noise was zero. For each testing session, we estimated the noise level and the mean/SD of the signal for every voxel. We used Monte Carlo simulation to obtain the noise ceiling. For each simulation, we generated a signal from the signal distribution, and generated a noisy data by adding the signal and the noise drawn from the noise distribution, and calculated the correlation between the signal and the data. We performed 1,000 simulations for each testing session, and took the median correlation as the noise ceiling. The ceiling was then averaged across sessions.

3.2.4 Human-face representations with encoding models and functional localizer

The ResNet-based encoding models were further used to simulate cortical representations of human faces, in comparison with the results obtained with a functional localizer applied to the same subjects. To simulate the cortical "face" representation, 2,000 human-face pictures were obtained by Google Image search. Each of these pictures was input to the voxel-wise encoding model, simulating a cortical response map as if it were generated when the subject was actually viewing the picture, as initially explored in previous studies[7, 30]. The simulated response maps were averaged across all the face pictures, synthesizing the cortical representation of human face as an object category.

To validate the model-synthesized "face" representation, a functional localizer[122] was used to experimentally map the cortical face areas on the same subjects. Each subject participated in three sessions of fMRI with a randomized block-design paradigm. The paradigm included alternating ON-OFF blocks with 12s per block. During each ON block, 15 pictures (12 novel and

3 repeated) from one of the three categories (face, object, and place) were shown for 0.5s per each picture with a 0.3s interval. The ON blocks were randomized and counter-balanced across the three categories. Following the same preprocessing as for the video-fMRI data, the block-design fMRI data were analyzed with a general linear model (GLM) with three predictors, i.e. face, object, and place. Cortical "face" areas were localized by testing the significance of a contrast (face>object and face > place) with $p<0.05$ and Bonferroni correction.

3.2.5 Synthesizing cortical representations of different categories

Beyond the proof of concept with human faces, the similar strategy was also extended to simulate the cortical representations of 80 categories through the ResNet-based encoding models. The category labels were shown in Fig. 3.3. These categories were mostly covered by the video clips used for training the encoding models. For each category, around 800 pictures were obtained by Google Image search with the corresponding label, and were visually inspected to replace any exemplar that belonged to more than one category. There was a total of 64,000 objects from 80 categories. The cortical representation of each category was generated by averaging the model-simulated response map given every exemplar within the category.

We focused on cortical representations of basic-level object categories, as opposed to individual images. Although the models were able to simulate and characterize cortical activations with each of the images, as already done in our prior study [7], herein the total number of images (64,000) was too large. This choice was also given our primary interest in representations of object knowledge, regardless of the luminance, position, and size of any object. However, the exclusive focus on category-average representations, may be biased by how categories were defined and how images were selected (by humans). More detailed analysis of responses to individual image exemplars is helpful to mitigate this bias or ambiguity[17].

3.2.6 Category selectivity

Following the above analysis, cortical representations were compared across categories to quantify the category selectivity of various locations and ROIs. For each voxel, its selectivity to category i against other categories i^c was quantified with equation (5), as previously suggested[123].

$$d_i' = \frac{\bar{r}_i - \bar{r}_{i^c}}{\sqrt{(\sigma_i^2 + \sigma_{i^c}^2)/2}} \qquad (5)$$

where $\bar{r}_i$ and σ_i^2 are the mean and variance of the responses to the exemplars in category i, and $\bar{r}_{i^c}$ and $\sigma_{i^c}^2$ were counterparts to all exemplars in other categories i^c. Irrespective of any specific category, the general category-selectivity for each voxel was its maximal d' index among all categories, i.e. $d' = \max_i\{d_i'\}$. A d' index of zero suggests non-selectivity to any category, and a higher d' index suggests higher category-selectivity. The category selectivity of any given voxel was also inspected by listing the categories in a descending order of their representations at the voxel. We also obtained the ROI-level category selectivity by averaging the voxel-wise selectivity across voxels and subjects. ROIs were defined in an existing cortical parcellation[62].

3.2.7 Categorical similarity and clustering in cortical representation

To reveal how the brain organizes categorical information, we assessed the similarity (i.e. the Pearson's correlation of the spatial response patterns across the predictable voxels with q<0.01 in permutation test and prediction accuracy r>0.2) in cortical representations between categories. Based on such inter-category similarity, individual categories were grouped into clusters using k-means clustering [124]. The goodness of clustering was measured as the modularity index, which quantified the inter-category similarities within the clusters relative to those regardless of the clusters[125]. The number of clusters was determined by maximizing the modularity index. To quantify the modularity index, the categorical similarity was viewed as a signed, weighted, and undirected network[125]. Each node represented one category, and each weighted edge represented the similarity between two categories. The modularity was then measured as the probability of having edges falling within clusters in the network against a random network (null case) with the same number of nodes and edges placed at random preserving the degree of each node. Specifically, given a *positive* weighted matrix $\boldsymbol{S}$ (S_{ij} denotes the weight between categories i and j, and $S = 2\sum_i\sum_j S_{ij}$ denotes the double total weight), the modularity index Q was defined as $Q = \sum_i\sum_j(p_{ij} - q_{ij})\delta(C_i, C_j)$, where $p_{ij} = S_{ij}/S$ is the probability of connecting category i and j in the network with edge weight S_{ij}, $q_{ij} = (\sum_j S_{ij}/S)(\sum_i S_{ij}/S)$ denotes the expected probability of having edge between i and j in random networks, and $\delta(C_i, C_j)$ is the Kronecker delta function with value 1 if i and j are in the same cluster and 0 otherwise. Since the correlation coefficients ranged from -1 to 1, we separated the positive and negative weights by $S_{ij} = S_{ij}^+ - S_{ij}^-$ where $S_{ij}^+ = \max\{0, S_{ij}\}$ and $S_{ij}^- = \max\{0, -S_{ij}\}$, and calculated their corresponding modularity

Q^+ and Q^-. Then the total modularity was quantified as $Q = \frac{s^+}{s^+ + s^-} Q^+ - \frac{s^-}{s^+ + s^-} Q^-$. The significance of the modularity index was assessed by permutation test against the null distribution obtained from shuffling the pair-wise similarities randomly for 100,000 times. The larger modularity means the larger deviation from the null case and the better differentiation between clusters. Noted that higher similarity within clusters and less similarity across clusters gives larger modularity.

The similarity in cortical representation between different categories was compared with their similarity in semantic meaning. Here, we explored three different models to measure the semantic similarity between categories. For the first model, the semantic similarity between categories was evaluated as the Leacock-Chodorow similarity (LCH)[126] between the corresponding labels based on their relationships defined in the WordNet[127] – a directed graph of words (as the nodes) and their *is-a* relationships (as the edges). Briefly, LCH computes the similarity (s) between two labels based on the shortest path (p) that connects the labels in the taxonomy and the maximum depth (d) of the taxonomy in which the labels occur through $s = -\log(p/2d)$. The second model was the word2vec model that represented text words in a continuous vector space that captured a large number of precise syntactic and semantic word relationship[128]. We used the published model that was pretrained by Google on 100 billion words from Google News. The model was trained to accurately predict surrounding words given the current word. We used it to transform the category labels to vectors, and then calculated the semantic similarity between labels as the cosine distance of their corresponding vectors. The third model was the GloVe model that also represented words in vectors and captured fine-grained semantic and syntactic regularities using vector arithmetic[129]. GloVe was trained on global word-word co-occurrence statistics from a corpus of text. Similarly, we used the pretrained GloVe (trained on a large corpus including 840 billion tokens) to derive the vectors of category labels and calculate their semantic similarity as the cosine distance between the vectors. After obtaining the inter-category semantic similarity (LCH, word2vec, or GloVe), we evaluated the Pearson's correlation between the cortical and semantic similarities. Before computing the correlation, the cortical similarity was transformed to z-score by using the Fisher's z-transformation. Since the similarity was symmetric, the correlation was computed over the values in the upper (or equivalently the lower) triangular region of the similarity matrix[130]. The significance was assessed by random permutation of the category labels (i.e. reordering rows and columns of the

cortical similarity matrix according to this permutation and computing the correlation). By repeating the permutation step 10,000 times, we obtained a distribution of correlations simulating the null hypothesis that the two similarity matrices are unrelated[130].

3.2.8 Layer-wise contribution to cortical categorical representation

We also asked which levels of visual information contributed to the clustered organization of categorical representations in the brain. To answer this question, the cortical representation of each category was dissected into multiple levels of representations, each of which was attributed to one single layer of features. For a given category, the features extracted from every exemplar of this category were kept only for one layer in the ResNet, while setting to zeros for all other layers. Through the above trained encoding models (see ***Encoding models*** in **Materials and Methods**), the single-layer visual features were projected onto a cortical map that only represented a certain level of visual information shared in the given category. The similarity and modularity in cortical representations of individual categories were then re-evaluated as a function of the layer in the ResNet. The layer with the highest modularity index contributed the most to the clustered organization in cortical categorical representation. The features encoded by this layer were visualized for more intuitive understanding of the types of visual information underlying the clustered organization. The feature visualization was based on an optimization-based technique[131]. Briefly, to visualize the feature encoded by a single unit in the ResNet, the input to the ResNet was optimized to iteratively maximize the output from this unit, starting from a Gaussian random pattern. Four optimized visualizations were obtained given different random initialization.

After obtained the layer-wise similarities in cortical representations of object categories, we further evaluated the correlation between the cortical similarity and the semantic similarity for each layer, and assessed its significance by using the aforementioned permutation test ($p=0.0001$).

3.2.9 Finer clustering of categorical representation

Considering object categories were defined hierarchically in semantics[127], we asked how hierarchy of categorization[99]. More specifically, we tested whether the representational similarity and distinction in a larger spatial scale gave rise to a coarser level of categorization, whereas the representation in a smaller spatial scale gave rise to a finer level of categorization. To

do so, we first examined the category representation in the scale of the entire visual cortex predictable by the encoding models, and clustered the categories into multiple clusters by using the clustering analysis of the representational similarity in this large scale. The resulting clusters of categories were compared with the superordinate-level semantic categories. Then, we focused on a finer spatial scale specific to the regions where category representations overlapped within each cluster. The cluster-specific region included the cortical locations whose activation was significantly higher for objects in the cluster compared to 50,000 random and non-selective objects ($p<0.01$, two-sample t-test, Bonferroni correction). Given the spatial similarity of category representation in this finer scale, we defined sub-clusters within each cluster using the same clustering analysis as for the large-scale representation. The sub-clusters of categories were compared and interpreted against semantic categories in a finer level.

3.3 Results

3.3.1 ResNet predicted widespread cortical responses to natural visual stimuli

In line with recent studies[7-10, 22, 30], we used a deep convolutional neural network to establish predictive models of cortical fMRI representations of natural visual stimuli. Specifically, we used ResNet – a deep residual network for computer vision[21]. With a much deeper architecture, ResNet offers more fine-grained layers of visual features, and it performs better in image recognition than similar but shallower networks, e.g. AlexNet[19] as explored in prior studies[7-10, 22, 30, 56]. In this study, we used ResNet to extract visual features from video stimuli, and used the extracted features to jointly predict the evoked fMRI response through a voxel-wise linear regression model. This encoding model was trained with a large amount of fMRI data during a training movie (12.8 hours for Subject 1, and 2.4 hours for Subject 2, 3), and tested with an independent testing movie (40 minutes).

The encoding accuracy (i.e. the correlation between the predicted and measured fMRI signals during the testing movie) was overall high ($r = 0.43\pm0.14$, 0.36 ± 0.12, and 0.37 ± 0.11 for Subject 1, 2 and 3, respectively) and statistically significant (permutation test, corrected at FDR $q<0.01$) throughout the visual cortex in every subject (Fig. 3.1.a). The encoding accuracy was comparable among the higher-order ventral-stream areas, e.g. fusiform face area [132] and parahippocampal place area (PPA), as well as early visual areas, e.g. V1, V2, and V3 (Fig. 3.1.c).

The accuracy was relatively lower at dorsal-stream areas such as lateral intraparietal area (LIP), frontal eye fields (FEF), parietal eye fields (PEF), but not the middle temporal area (MT) (Fig. 3.1.c). Different cortical regions were preferentially correlated with distinct layers in ResNet. The lower to higher level visual features encoded in ResNet were gradually mapped onto areas from the striate to extrastriate cortex along both ventral and dorsal streams (Fig. 3.1.b), in agreement with previous studies[7, 22, 30, 53, 56, 133]. The prediction accuracy was consistently higher with (the deeper) ResNet than with (the shallower) AlexNet (Fig. 3.1.c). These results suggest that the ResNet-based voxel-wise encoding models offer generalizable computational accounts for the complex and nonlinear relationships between natural visual stimuli and cortical responses at widespread areas involved in various levels of visual processing.

3.3.2 Encoding models predicted cortical representations of various object categories

As explored before[7, 30], the voxel-wise encoding models constituted a high-throughput platform to synthesize cortical activations with an infinitely large number of natural pictures that are unrealistic or expensive to acquire with most experimental approaches. Here, we used this strategy to predict the pattern of cortical activation with each of the 64,000 natural pictures from 80 categories with on average 800 exemplars per category. By averaging the predicted activation maps across all exemplars of each category, the common cortical activation within this category was obtained to report its cortical representation.

For example, averaging the predicted responses to various human faces revealed the cortical representation of the category “face” regardless of the position, size, color, angle, perspective of various faces (Fig. 3.2.a). Such a model-simulated “face” representation was consistent with the fMRI-mapping result obtained with a block-design functional localizer that contrasted face vs. non-face pictures (Fig. 3.2.b). In a similar manner, cortical representations of all 80 categories were individually mapped (Fig. 3.3). The resulting category representations were found not only along the ventral stream, but also along the dorsal stream albeit with relatively lower amplitudes and a smaller extent.

For each voxel, the model-predicted response as a function of category was regarded as the voxel-wise profile of categorical representation. The category selectivity – a measure of how a voxel was selectively responsive to one category relative to others[123], varied considerably across cortical locations (Fig. 3.4.a). Voxels with higher category selectivity were clustered into discrete

regions including the bilateral PPA, FFA, lateral occipital (LO) area, the temporo-parietal junction (TPJ), as well as the right superior temporal sulcus (STS) (Fig. 3.4.a). The profile of categorical representation listed in a descending order (Fig. 3.4.b), showed that FFA, OFA, and pSTS were selective to humans or animals (e.g. man, woman, monkey, cat, lion); PPA was highly selective to places (e.g. kitchen, office, living room, corridor); and the ventral visual complex (VVC) was selective to man-made objects (e.g. cellphone, tool, bowl, car). In general, the ventral stream tended to be more category-selective than early visual areas (e.g. V1, V2, V3) and dorsal-stream areas (e.g. MT, LIP) (Fig. 3.4.c).

3.3.3 Distributed, overlapping, and clustered representations of categories

Although some ventral-stream areas (e.g. PPA and FFA) were highly (but not exclusively) selective to a certain category, no category was represented by any single region alone (Fig. 3.3). As suggested previously[80], object categories were represented distinctly by distributed but partially overlapping networks [27]. In the scale of the nearly entire visual cortex that was predictable by the encoding models (Fig. 3.1.a), the spatial correlations in cortical representation between different categories were shown as a representational similarity matrix (Fig 3.5.a). This matrix revealed a clustered organization: categories were clustered into three groups such that cortical representations were more correlated among categories within the same group than across different groups (Fig. 3.5.a, left), and the degree of clustering (quantified as the modularity index, Q) was high (Q=0.35). Interestingly, categories clustered together on the basis of their cortical representations tended to have higher conceptual similarities, or closer relationships between the corresponding category labels as measured by their Leacock-Chodorow (LCH) similarity in WordNet[126] (Fig. 3.5.a, middle), or by the cosine distance between their vector representations after word2vec[128] or GloVe[129] transformation [27]. Regardless of the distinct methods for measuring the semantic similarity, there was a significant correlation between the similarity in cortical representation and the similarity in semantics across all pairs of categories (Fig. 3.5.a, right). Moreover, we examined the category representations in a finer scale confined to individual visual areas (V1, V2, V3, LO, FFA, PPA). For each of these areas, we evaluated the correlation between representational similarity and semantic similarity across all pairs of categories. The correlation tended to increase from lower (e.g. V1) to higher (e.g. FFA/PPA) areas in the ventral stream [27]. However, the correlation was significant ($p<0.0001$, permutation test) not only in

higher ventral-stream areas, but also in mid-level areas (e.g. LO) or even lower areas (V2, V3). In sum, categories with closer cortical representations tend to bear similar semantic meanings, in the spatial scale of the whole brain as well as visual areas at different stages of visual processing.

The representational clusters in the entire visual cortex grouped basic-level categories into super-ordinate-level categories. The first cluster included non-biological objects, e.g. airplane, bottle and chair; the second cluster included biological objects, e.g. humans, animals, and plants; the third cluster included places and scenes (e.g. beach, bedroom) (Fig. 3.5.b). The cortical representation averaged within each cluster revealed the general cortical representations of superordinate categories. As shown in Fig. 3.5.b, non-biological objects were represented by activations in bilateral sub-regions of the ventral temporo-occipital cortex (e.g. VVC); biological objects were represented by activations in the lateral occipital cortex and part of the inferior temporal cortex (e.g. FFA) but deactivations in parahippocampal cortex (e.g. PPA); background scenes were represented by activations in PPA but deactivations in the lateral occipital complex, partly anti-correlated with the activations with biological objects. The spatial correlations between the cortical representations of biological objects and background scenes were on average –0.17±0.29, which should be cautiously taken as a tendency of anti-correlation instead of strong evidence for precisely opposite patterns of representations of these two kinds of categories.

3.3.4 Mid-level visual features primarily accounted for superordinate categorization

Which levels of visual features accounted for such a clustered organization of cortical category representation? To address this question, we simulated the cortical representation of single-layer features of every image exemplar in each category, by setting to zero all other layers in ResNet except one before inputting the feature representations into voxel-wise linear encoding models. Then we evaluated the similarity in cortical representation between categories at an increasing level of visual processing, progressively defined by the first through last layer in ResNet. Fig. 3.6.a (left) shows the representational similarity matrix attributed to features in each layer, thus decomposing the clustered organization in Fig. 3.5.a by layers. In the earliest level of visual processing as specified by V1-like neurons in the first layer of ResNet, the similarity (or dissimilarity) among different categories was not apparent within (or across) the three superordinate categories (non-biological objects, biological objects, and background scenes). At layer 4, non-biological objects differed themselves from biological objects or background scenes,

as the representational similarity appeared to reveal two clusters, rather than three clusters. Starting from layer 10 through 19, the three clusters emerged in the corresponding representational similarity matrices. Starting from layer 25, anti-correlations became clearly notable between the cluster of biological objects and the cluster of background scenes.

In a more quantitative way, we evaluated the modularity index of the three-cluster organization due to layer-wise features. Fig. 3.6.a (right) shows the modularity index as a function of the layer in ResNet. It suggests that the clustering of basic-level categories into superordinate categories emerged progressively and occurred in many levels of visual processing, while the clustering was the most apparent in the middle level (i.e. layer 31 in ResNet). To gain intuition about the types of visual information from the 31st layer, the features encoded by individual units in this layer were visualized. Fig. 3.6.b illustrates the visualizations of some example features, showing shapes or patterns (both 2-D and 3-D), animal or facial parts (e.g. head and eye), scenic parts (e.g. house and mountain). Beyond these examples, other features were of similar types.

In addition, we evaluated the correlation between the inter-category semantic similarity and the corresponding similarity in cortical representation of the features in each layer. It turned out that the layer-wise correlations were significant ($p<0.001$) for middle and high-level features, and the greatest correlation was not necessarily in the highest layer, but in the middle layer (around layer 31) (Fig. 3.6c). It suggests that semantic relationships emerge from object attributes in different levels of visual processing, and that the mid-level attributes (e.g. object shapes or parts) contribute the most to superordinate-level categorization.

3.3.5 Clustered organization of cortical representation within superordinate categories

We further asked whether the similarly clustered organization could be extended to a lower level of categorization. That is, whether object representations were organized into sub-clusters within each superordinate-level cluster. For this purpose, we confined the scope of analysis from the whole visual cortex to finer spatial scales highlighted by the co-activation patterns within biological objects, non-biological objects, or background scenes (Fig. 3.7.a). For example, within the regions where biological objects were represented (Fig. 3.7.a, top), the representational patterns were further clustered into four sub-clusters: terrestrial animals, aquatic animals, plants, and humans (Fig. 3.7.b, top). Similarly, the fine-scale representational patterns of background scenes were clustered into two sub-clusters corresponding to artificial (e.g. bedroom, bridge, restaurant)

and natural scenes (e.g. falls, forest, beach) (Fig. 3.7, middle). However, the two clusters of non-biological objects did not bear any reasonable conceptual distinction (Fig. 3.7, bottom).

We also evaluated the layer-wise contribution of visual features to the fine-scale representational similarity and clustering. For biological objects, the modularity index generally increased from the lower to higher layer, reaching the maximum at the highest layer (Fig. 3.8.a). Note that the highest layer encoded the most abstract and semantically relevant features, whose visualizations revealed the entire objects or scenes [27] rather than object or scenic parts (Fig. 3.6.b). In contrast, the modularity index reached the maximum at the 28th layer for background scenes (Fig. 3.8.b), but was relatively weak and less layer-dependent for non-biological objects (Fig. 3.8.c).

3.4 Discussion

This study demonstrates a high-throughput computational strategy to characterize hierarchical, distributed, and overlapping cortical representations of visual objects and categories. Results suggest that information about visual-object category entails multiple levels and domains of features represented by distributed cortical patterns in both ventral and dorsal pathways. Categories with similar cortical representations are more semantically related to one another. In a large scale of the entire visual cortex, cortical representations of objects are clustered into three superordinate categories (biological objects, non-biological objects, and background scenes). In a finer spatial scale that is specific to each cluster, cortical representations are organized into sub-clusters for finer categorization, e.g. biological objects are categorized into terrestrial animals, aquatic animals, plants, and humans. The clustered organization of cortical representation is more observable for object features in middle and high levels of complexity compared to low-level features. Therefore, the brain categorizes visual objects through the hierarchically clustered organization of object attributes emerging from various levels of visual processing, rather than any operation that only occurs at the highest level of the ventral-stream hierarchy.

Central to this study is the use of the categorization-driven deep ResNet for synthesizing the cortical representations of thousands of natural visual objects from many categories. This strategy has a much higher throughput in sampling a virtually infinite number of exemplars of visual objects[7, 30], compared to prior studies that are limited to fewer categories with much fewer exemplars per category[77, 79, 111-113]. The sample size could be further extendable, since

the ResNet-based encoding models presumably account for the relationships between cortical responses and visual features that are generalizable to different and new natural images, objects, and categories which the models have not been explicitly trained with. The model predictions are highly accurate and consistent with experimentally observed cortical responses to video stimuli and cortical representations to specific objects (e.g. human faces). The encoding accuracy may be further improved given an even larger and more diverse video-fMRI dataset to train the model, and a more biologically relevant deep neural net that better matches the brain and better performs in computer-vision tasks[9]. In this sense, the encoding models in this study are based on so far the largest video-fMRI training data from single subjects; and ResNet also outperforms AlexNet in categorizing images[19, 21] and predicting the brain (Fig. 3.1.c). The encoding models reported here are thus arguably more powerful for predicting and mapping hierarchical cortical representations in the entire visual cortex, compared to other conceptually similar models in prior studies[7-10, 22, 30].

What is also advantageous is that ResNet decomposes category information into multiple layers of features progressively emerging from low to middle to high levels. As such, ResNet offers a computational account of hierarchical cortical processing for categorization, yielding quantitative description of every object or category in terms of different layers of visual features. Mapping the layer-wise features from the ResNet onto the brain helps to address what drives the cortical organization of object knowledge and supports various levels of categorization.

The ResNet is trained with large-scale image set (~1.3 million natural images) for recognizing 1,000 visual object categories[21]. Though specific categories are used in training the ResNet, the trained model is generalizable to represent the semantics in our training and testing stimuli, and is transferrable for recognizing new categories based on the generic representations in the learned feature space for transfer learning[50, 134]. The generalizability of the feature space allows for prediction of the cortical representations of a wide range of categories far beyond those that the network has been explicitly trained. For example, the model is able to predict the face representation even though the ResNet is not trained for recognizing faces (Fig. 3.2).

Our results support the notion that visual-object categories are represented by distributed and overlapping cortical patterns[80] rather than clustered regions[57, 114, 115]. Given this notion, the brain represents a category not as a single entity but a set of defining attributes that span multiple domains and levels of object knowledge. Different objects bear overlapping

representational patterns that are both separable and associable, allowing them to be recognized as one category in a particular level, but as different categories in another level. For example, a lion and a shark are both animals but can be more specifically categorized as terrestrial and aquatic animals, respectively. The distributed and overlapping object representations, as weighted spatial patterns of attribute-based representations[105], constitute an essential principle underlying the brain's capacity for multi-level categorization.

Category representations may become highly selective at spatially clustered regions[57, 114, 115]. The category-selective regions are mostly in the ventral temporal cortex (Fig. 3.4), e.g. the FFA, PPA, and LO. The existence of category-selective regions does not contradict with distributed category representation. Instead, a region specific to a given category is thought to emerge from its connectivity with other locations that represent the defining attributes of that category[135], or subserve the category-specific action and cognition[136].

The cortical representational similarity between different categories is highly correlated with their semantic relationship (Fig. 3.5). In other words, the semantic relationship is preserved by cortical representation. This finding lends support for the notion of a continuous semantic space underlying the brain's category representation[60], which is a parsimonious hypothesis to bridge neural representation and linguistic taxonomy[87]. However, category information is not limited to semantic features, but includes hierarchically organized attributes, all of which define categories and their conceptual relationships. For example, "face" is not an isolated concept; it entails facial features ("eyes", "nose", "mouth"), each also having its own defining features. The similarity and distinction between categories may be attributable to one or multiple levels of features. In prior studies[60], the hierarchical nature of category information is not considered as every exemplar of each category is annotated by a pre-defined label. This causes an incomplete account of category representation, leaving it difficult to disentangle the various levels of category information that may be used to associate or distinguish categories.

We have overcome this limit by disentangling multiple layers of features from visual objects and evaluating their respective cortical representations. Our results show that different levels of features make distinctive contributions to the clustering of category representation in the visual cortex. Coarse categories (i.e. biological objects, non-biological objects, and background scenes) are most attributable to mid-level features, e.g. shapes, textures, and object parts (Fig. 3.6). In a finer level of categorization, terrestrial animals, aquatic animals, plants, and humans are most

distinguishable in the semantic space; categorization of man-made and natural scenes is most supported by mid-level features (Fig. 3.8). In addition, the semantic similarity between categories is correlated with the spatial similarity in cortical representation of their middle to high-level visual features (Fig. 3.6), not necessarily confined to one level or domain of features or a single cortical region, e.g. ITC[117]. Recent studies have also shown that the cortical organization of visual objects may be explained in part by similarity in low-level visual features[137-139], shape[55, 138, 140-144], and the real-word or conceptual size of objects[145, 146]. This study further expands the dimension of visual or conceptual features beyond what can be intuitively defined[147], by using data-driven features extracted from ResNet[21],

This study is focused on the use of CNN-based encoding models to study the brain's mechanism for categorization, rather than only on the validation of a CNN against neuroscience data. Arguably, if a model is able to predict cortical responses to natural visual stimuli, it is reasonable to use the model as a computational tool to characterize the brain itself. Similar ideas have been utilized to map the brain's semantic representation by using semantics-based encoding models[60], yielding insightful findings about how the brain represents natural language. However, it should be noted that although it is successful explaining significant variance of cortical responses to video stimuli, ResNet is not a perfect model of the visual cortex, and does not reach the noise ceiling. ResNet, or other types of feed-forward-only CNN, ignores the temporal relationships between video frames. Thus, the ResNet-based encoding models are more suitable to be trained with well-separated static image stimuli, which would take much longer time to acquire an equivalent amount of training data (as with video stimuli) for training the encoding models with millions of parameters. In addition, ResNet does not include any feedback connections or account for active attention, and fails to mimic the brain's ability of unsupervised learning[148]. For these reasons, ResNet is by no means an ultimate model of the visual cortex. Nevertheless, a feed-forward CNN is appropriate for modeling the brain's mechanism for rapid visual categorization, which is arguably mostly feed-forward[89, 96]. Our results suggest that CNN can be used to reproduce the cortical organization of category representations, selectivity, and clustering, which often require extensive experimental efforts to reveal[77, 79, 111-113]. The CNN-based encoding models may allow researchers (or students) to run "virtual-fMRI" experiments with arbitrary visual stimuli, simulate cortical activations, and accordingly raise hypotheses for testing with real experiments. In the meantime, it awaits future studies to validate this strategy with more

experimental data and a rich stimulus set with different configurations, and to develop more biologically plausible models to replace CNN in this computational strategy.

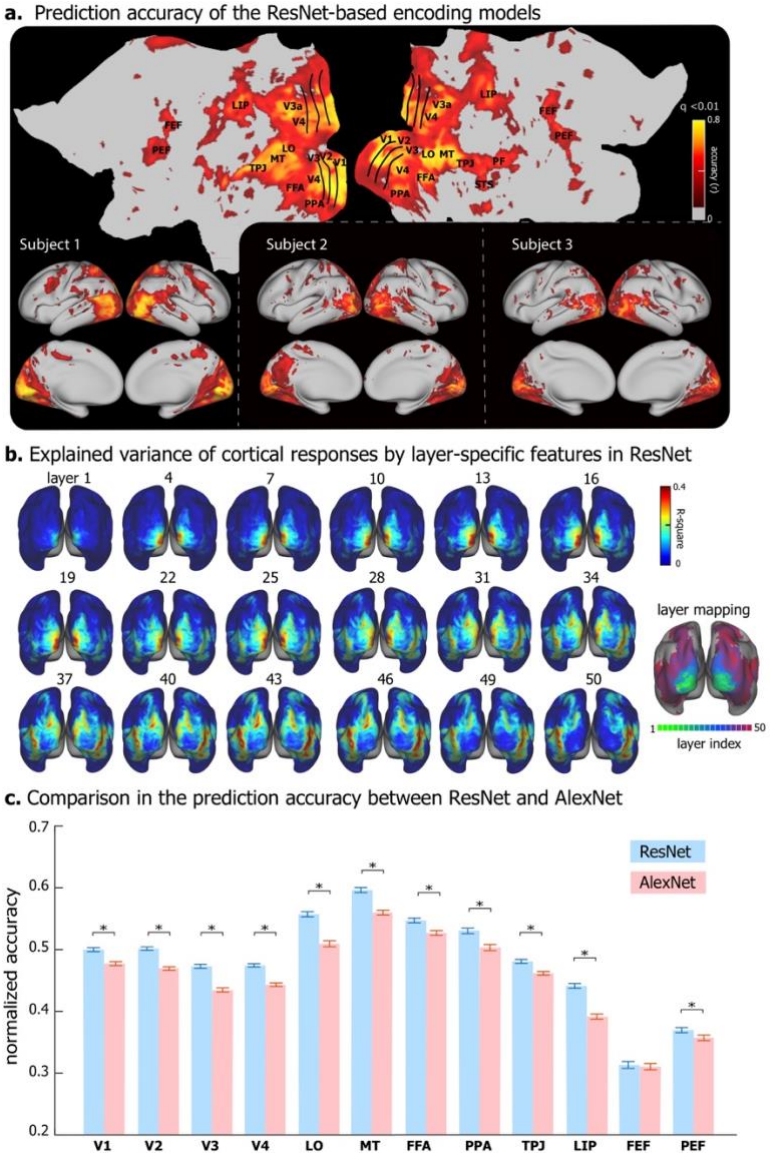

Figure 3.1 **DNN-based Voxel-wise encoding models.** (a) Performance of ResNet-based encoding models in predicting the cortical responses to novel testing movies for three subjects. The accuracy is measured by the average Pearson's correlation coefficient (r) between the predicted and the observed fMRI responses across five testing movies ($q<0.01$ after correction for multiple testing using the false discovery rate (FDR) method, and with threshold $r>0.2$). The prediction accuracy is displayed on both flat (top) and inflated (bottom left) cortical surfaces for Subject 1. (b) Explained variance of the cortical response to testing movie by the layer-specific visual features in ResNet. The right shows the index to the ResNet layer that most explains the cortical response at every voxel. **(c)** Comparison between the ResNet-based and the AlexNet-based encoding models. Each bar represents the mean±SE of the prediction accuracy (normalized by the noise ceiling, i.e. dividing prediction accuracy (r) by the noise ceiling at every voxel) within a ROI across voxels and subjects, and * represents a significance p-value ($p<0.001$) with paired t-test.

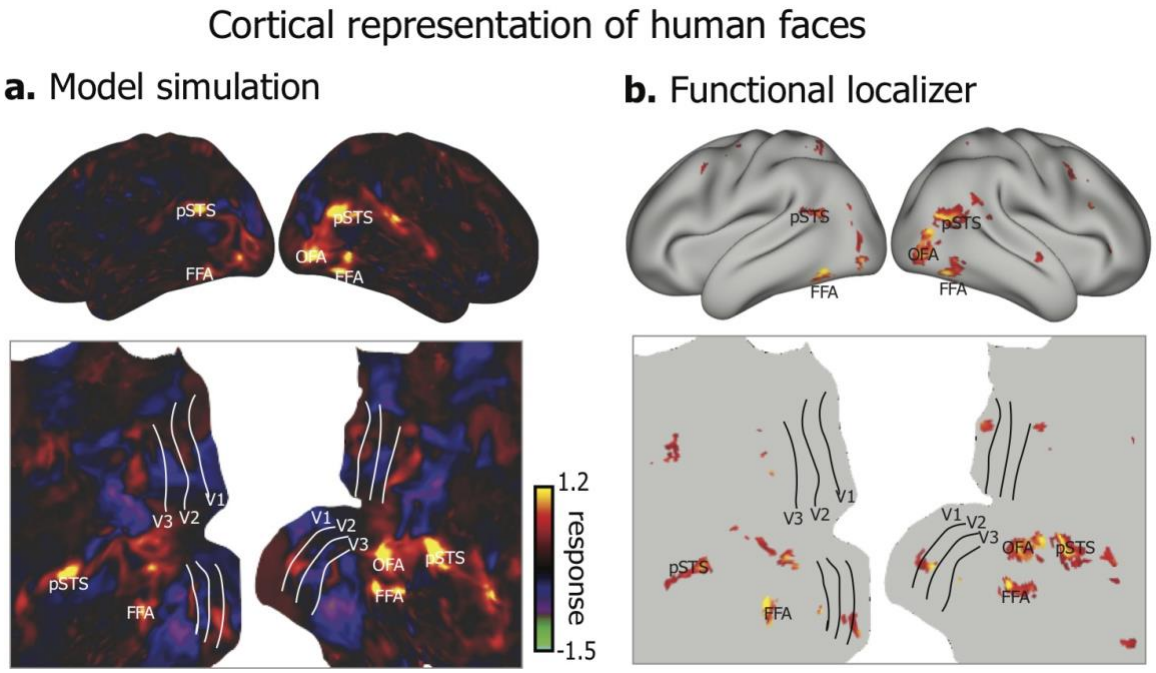

Figure 3.2 **Human-face representations with encoding models and functional localizer.** (a) Model-simulated representation of human face from ResNet-based encoding models. The representation is displayed on both inflated (top) and flat (bottom) cortical surfaces. **(b)** Localizer activation maps comprising regions selective for human faces, including occipital face area (OFA), fusiform face area [132], and posterior superior temporal sulcus (pSTS).

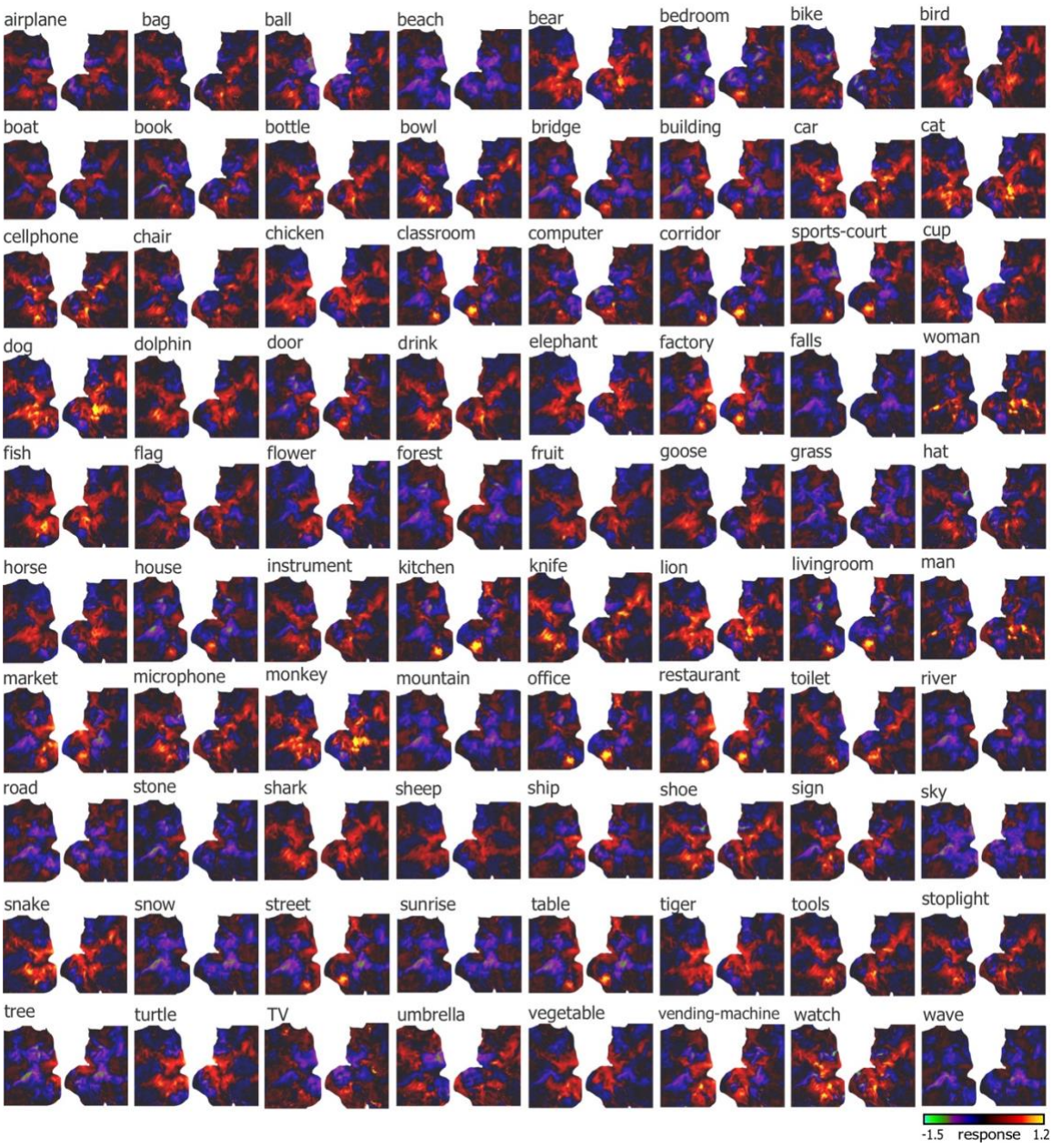

Figure 3.3 **Cortical representations of 80 object categories.** Each panel shows the representation map of an object category on flat cortical surface from Subject 1. The category label is on top left. The color bar shows the cortical response.

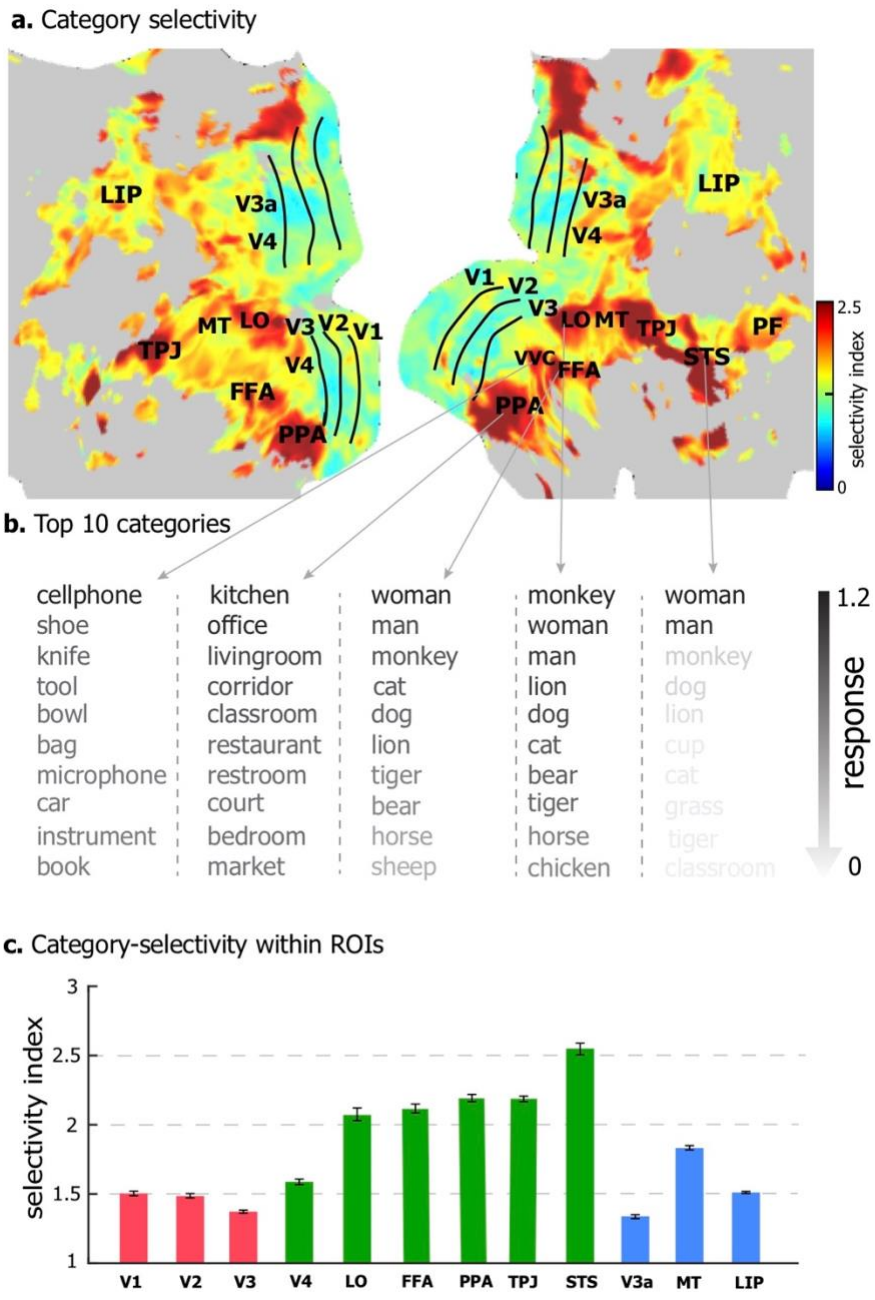

Figure 3.4 **Category-selectivity at individual cortical locations.** (a) The category-selectivity across the cortical surface. (b) The category-selectivity profile of example cortical locations. For each location, top 10 categories with the highest responses are showed in descending order. (c) Category-selectivity within ROIs (mean±SE) in the early visual areas (red), ventral stream areas [149], and dorsal stream areas (blue).

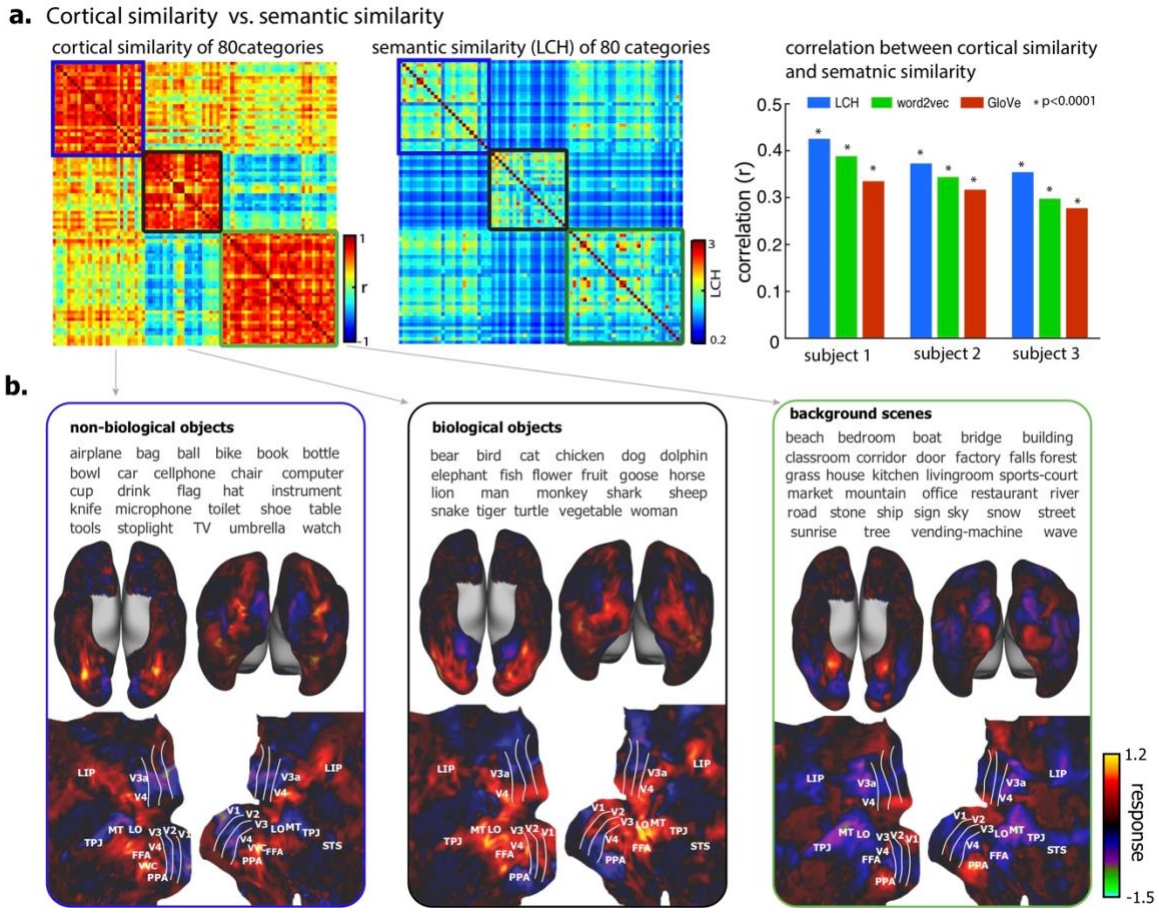

Figure 3.5 **Categorical similarity and clustering in cortical representation at the scale of the entire visual cortex.** (a) The left is the similarity matrix (Pearson's correlation r) of the cortical representations between categories. Each element represents the average cortical similarity between a pair of categories across subjects (see individual results in Supplementary Fig. S2 in [27]. It is well separated into three clusters with modularity Q=0.35. The middle is the similarity matrix of the semantic content between categories (measured by LCH). The right is the Pearson's correlation between the inter-category cortical similarity and the inter-category semantic similarity (with three different measures, i.e. the LCH similarity, the word2vec similarity, and the GloVe similarity). (b) These three clusters are related to three superordinate-level categories: non-biological objects, biological objects, and background scenes. The average cortical representations across categories within clusters are showed in the bottom on both inflated and flat cortical surfaces.

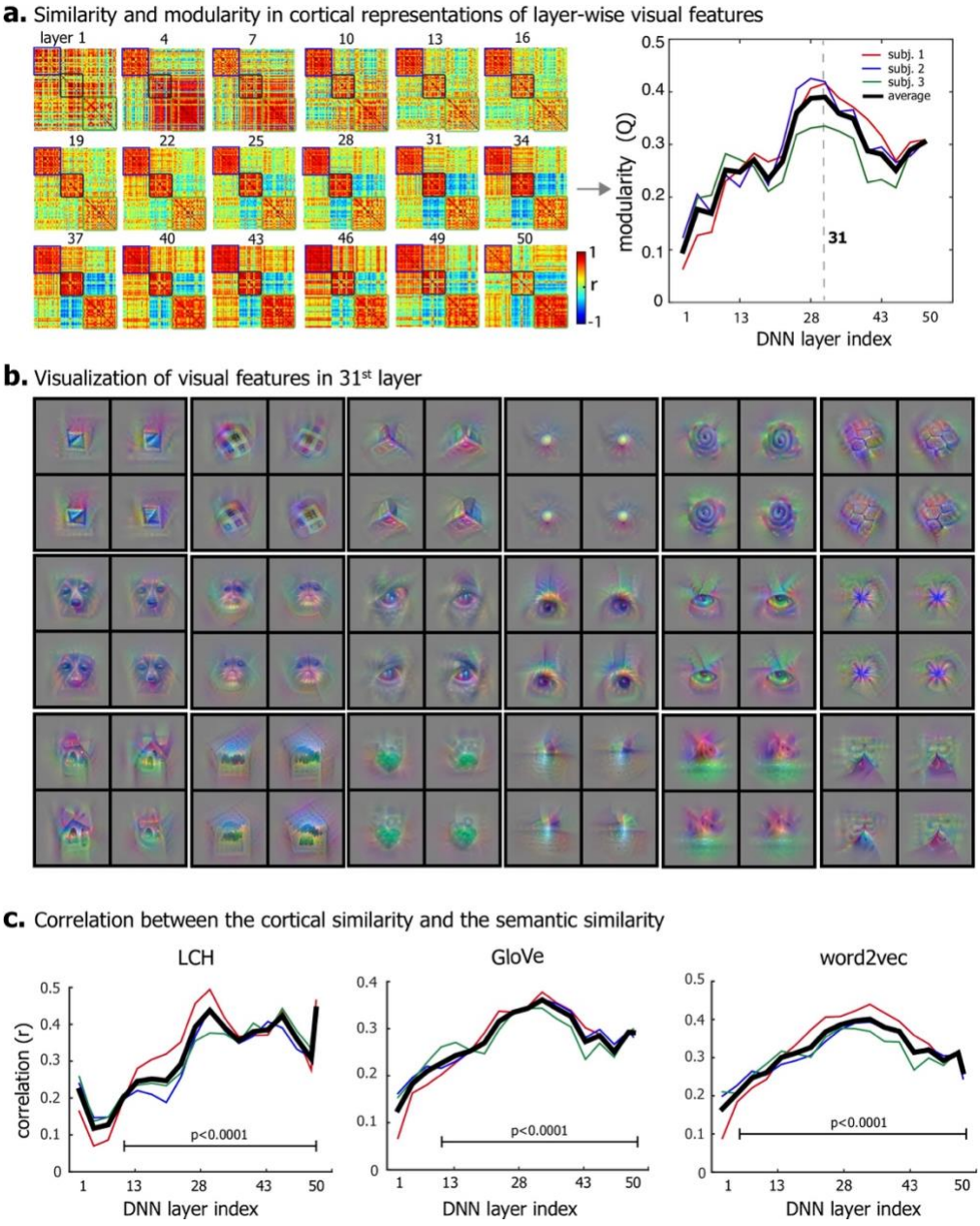

Figure 3.6 **Contribution of layer-wise visual features to the similarity and modularity in cortical representation.** (a) The left shows the similarity between categories in the cortical representations that are contributed by separated category information from individual layers. The order of categories is the same as in Figure 3.6.a. The right plot shows the modularity index across all layers. The visual features at the middle layers have the highest modularity. (b) 18 example visual features at the 31st layer are visualized in pixel space. Each visual feature shows 4 exemplars that maximize the feature representation. (c) The correlation between the inter-category cortical similarity across layers and the inter-category semantic similarity (with three different measures, i.e. the LCH similarity, the word2vec similarity, and the GloVe similarity).

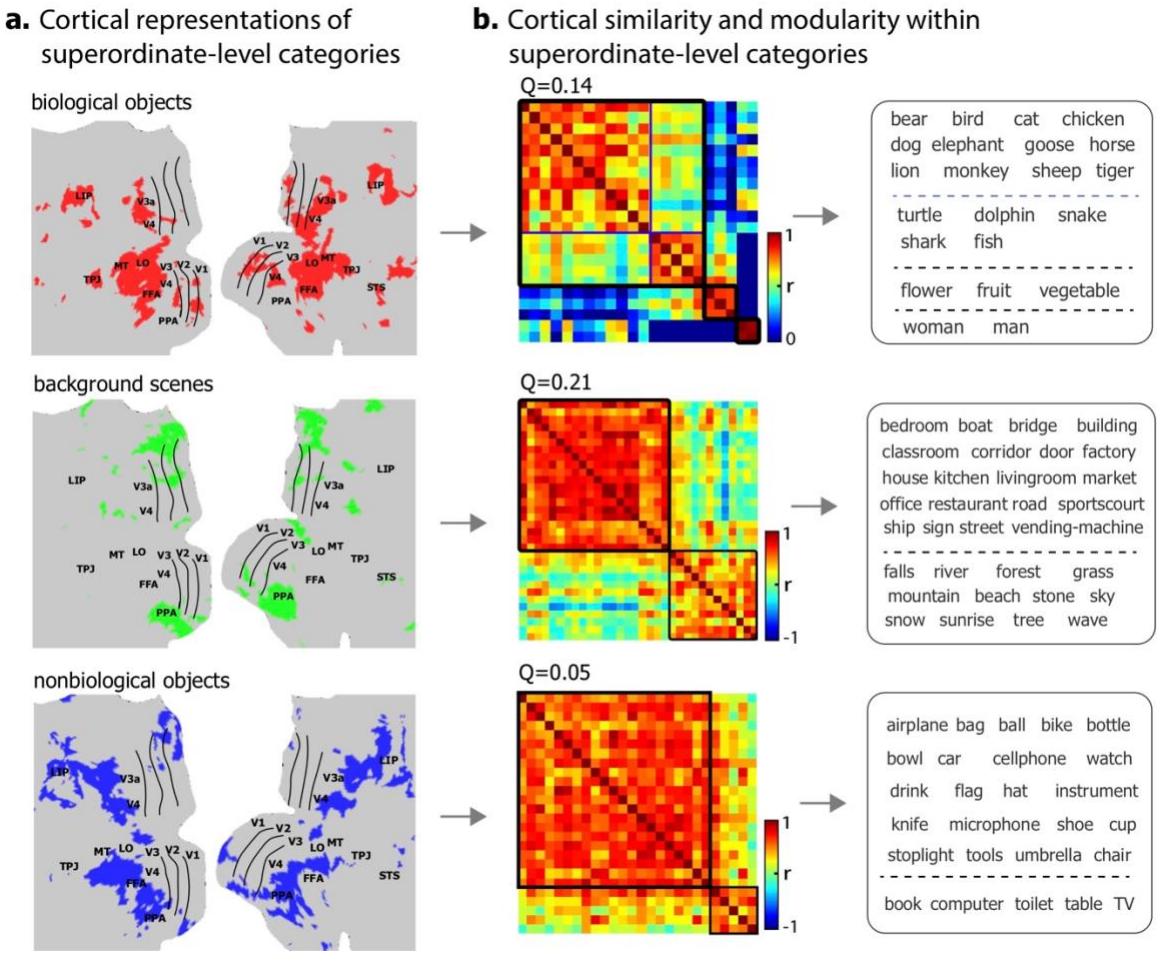

Figure 3.7 **Categorical similarity and clustering in cortical representation within superordinate-level categories.** (a) Fine-scale cortical areas specific to each superordinate-level category: biological objects (red), background scenes [149] and non-biological objects (blue). (b) The cortical similarity between categories in fine-scale cortical representation. The categories in each sub-cluster were displayed on the right. See individual results in Supplementary Fig. S2 in [27].

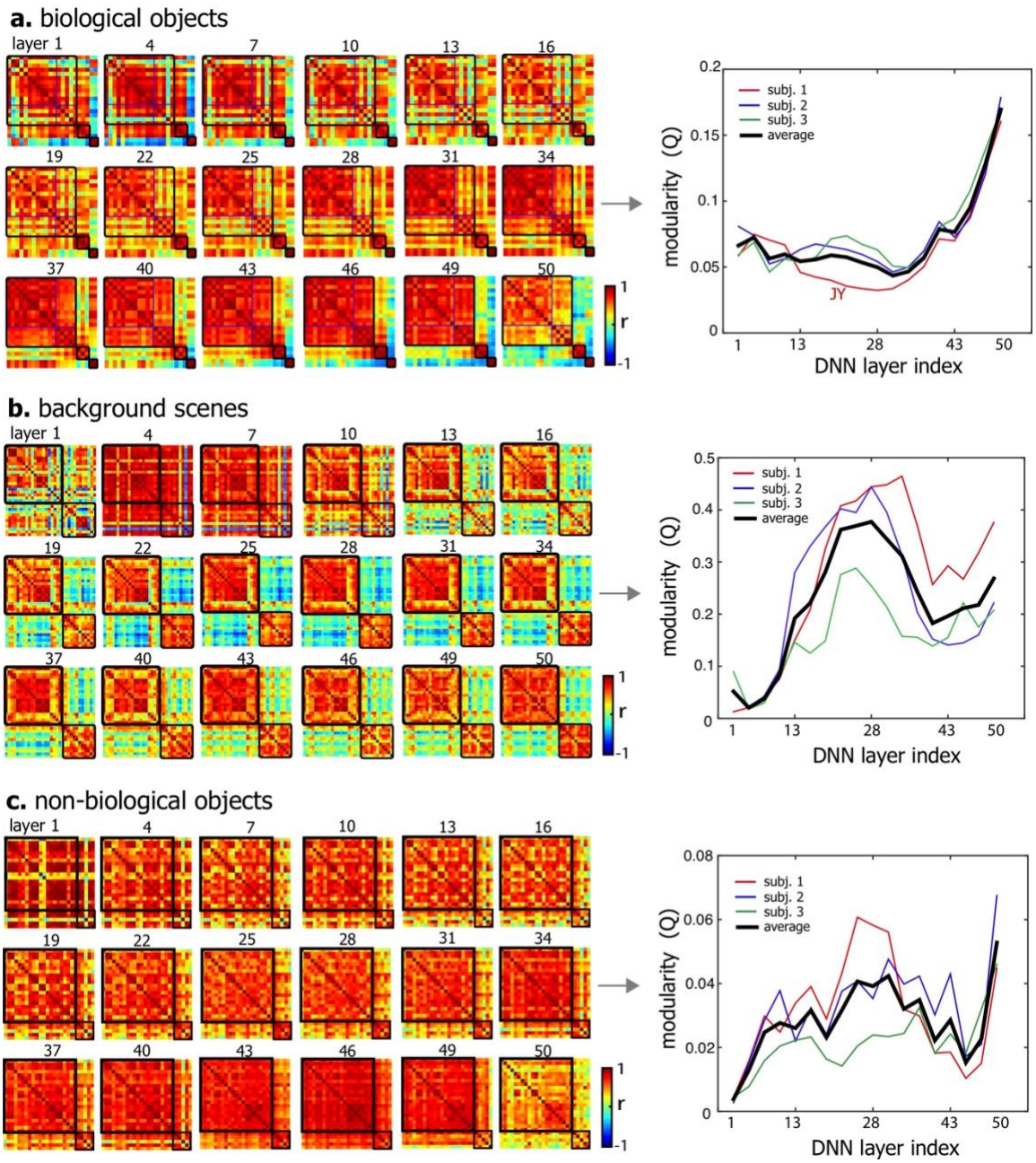

Figure 3.8 **Contribution of layer-wise visual features to the similarity and modularity in cortical representations within superordinate-level categories.** The left shows the similarity between categories in fine-scale cortical representations that are contributed by separated category information from individual layers. The order of categories is the same as in Figure 3.7. The right plot shows the modularity index across all layers. The highest-layer visual features show the highest modularity for biological objects.

4. TRANSFERRING AND GENERALIZING DEEP-LEARNING-BASED NEURAL ENCODING MODELS ACROSS SUBJECTS

*Modified and formatted for dissertation from the article in revision in *NeuroImage* [28]

4.1 Introduction

An important area in computational neuroscience is developing encoding models to explain brain responses given sensory input [150]. In vision, encoding models that account for the complex and nonlinear relationships between natural visual inputs and evoked neural responses can shed light on how the brain organizes and processes visual information through neural circuits [1, 3, 12, 151, 152]. Existing models may vary in the extent to which they explain brain responses to natural visual stimuli. For example, Gabor filters or their variations explain the neural responses in the primary visual cortex but not much beyond it [24, 26]. Visual semantics explain the responses in the ventral temporal cortex but not at lower visual areas [25, 60]. On the other hand, brain-inspired deep neural networks (DNN) [4], mimic the feedforward computation along the visual hierarchy [2, 3, 153, 154], match human performance in image recognition [6, 19, 20], and explain cortical activity over nearly the entire visual cortex in response to natural visual stimuli [7-9, 27, 29, 30, 35, 53, 155].

These models also vary in their complexity. In general, a model that explains brain activity in natural vision tends to extract a large number of visual features given the diversity of the visual world and the complexity of neural circuits. For DNN, the feature space usually has a very large dimension in the order of millions [6, 19-21]. Even if the model and the brain share the same representations up to linear transform [2], matching such millions of features onto billions of neurons or tens of thousands of neuroimaging voxels requires substantial data to sufficiently sample the feature space and reliably train the transformation from the feature model to the brain. For this reason, current studies have focused on only few subjects while training subject-specific encoding models with neural responses observed from each subject given hundreds to thousands of natural pictures [8, 30, 155], or several to tens of hours of natural videos [7, 27, 53]. However, a small subject pool incurs concerns on the generality of the conclusions drawn from such studies. Large data from single subjects are rarely available and difficult to collect especially for patients

and children. It is thus desirable to transfer encoding models across subjects to mitigate the need for a large amount of training data from single subjects.

Transferring encoding models from one subject to another should be feasible if different subjects share similar cortical representations of visual information. Indeed, different subjects show similar brain responses to the same natural visual stimuli [51, 52], after their brains are aligned anatomically. The consistency across subjects may be further improved by functional alignment of fine-grained response patterns [100, 156]. Recent studies have also shown that encoding [7, 8] or decoding [7, 157] models trained for one subject could be directly applied to another subject for reasonable encoding and decoding accuracies. Whereas these findings support the feasibility of transferring encoding and decoding models from one subject to another, it is desirable to consider and capture the individual variations in functional representations. Otherwise, the encoding and decoding performance is notably lower when the models are trained and tested for different subjects than for the same subject [7].

Beyond the level of single subjects, what is also lacking is a method to train encoding models for a group by using data from different subjects in the group. This need rises in the context of "big data", as data sharing is increasingly expected and executed [158-161]. For a group of subjects, combining data across subjects can yield much more training data than are attainable from a single subject. A population-wise encoding model also sets the baseline for identifying any individualized difference within a population. However, training such models with a very large and growing dataset as a whole is computationally inefficient or even intractable with the computing facilities available to most researchers [162].

Here, we developed methods to train DNN-based encoding models for single subjects or multiple subjects as a group. Our aims were to 1) mitigate the need for a large training dataset for each subject, and 2) efficiently train models with big and growing data combined across subjects. To achieve the first aim, we used pre-trained encoding models as the prior models in a new subject, reducing the demand for collecting extensive data from the subject in order to train the subject-specific models. To achieve the second aim, we used online learning algorithm [163] to adjust an existing encoding model with new data to avoid retraining the model from scratch with the whole dataset. To further leverage both strategies, we employed functional hyper-alignment [164] between subjects before transferring encoding models across subjects. Using experimental data for testing, we showed the merits of these methods in training the DNN-based encoding models to

predict functional magnetic resonance imaging (fMRI) responses to natural movie stimuli in both individual and group levels.

4.2 Methods and Materials

4.2.1 Experimental data

In this study, we used the video-fMRI data from our previous studies [7, 27]. The fMRI data were acquired from three human subjects (Subject JY, XL, and XF, all female, age: 22–25, normal vision) when watching natural videos. The videos covered diverse visual content representative of real-life visual experience.

For each subject, the video-fMRI data was split into three independent datasets for 1) functional alignment between subjects, 2) training the encoding models, and 3) testing the trained models. The corresponding videos used for each of the above purposes were combined and referred to as the "alignment" movie, the "training" movie, and the "testing" movie, respectively. For Subjects XL and XF, the alignment movie was 16 minutes; the training movie was 2.13 hours; the testing movie was 40 minutes. To each subject, the alignment and training movies were presented twice, and the testing movie was presented ten times. For Subject JY, all the movies for Subjects XL and XF were used; in addition, the training movie also included 10.4 hours of new videos not seen by Subjects XL and XF, which were presented only once.

Despite their different purposes, these movies were all split into 8-min segments, each of which was used as continuous visual stimuli during one session of fMRI acquisition. The stimuli ($20.3^{o}\times20.3^{o}$) were delivered via a binocular goggle in a 3-T MRI system. The fMRI data were acquired with 3.5 mm isotropic resolution and 2 s repetition time, while subjects were watching the movie with eyes fixating at a central cross. Structural MRI data with T_1 and T_2 weighted contrast were also acquired with 1 mm isotropic resolution for every subject. The fMRI data were preprocessed and co-registered onto a standard cortical surface template [49]. More details about the stimuli, data acquisition and preprocessing are described in our previous papers [7, 27, 29].

4.2.2 Nonlinear feature model based on deep neural network

The encoding models took visual stimuli as the input, and output the stimulus-evoked cortical responses. As shown in Fig. 4.1, it included two steps. The first step was a nonlinear feature model, converting the visual input to its feature representations; the second step was a voxel-wise

linear response model, projecting the feature representations onto the response at each fMRI voxel [7, 8, 24-27, 29, 30, 35, 53, 60, 155]. The feature model is described in this sub-section, and the response model is described in the next sub-section.

In line with previous studies [8, 53] [7, 27, 30, 155], a deep neural network (DNN) was used in the present study as the feature model to extract hierarchical features from visual input. Here, a specific version of the DNN, i.e. deep residual network (ResNet) [21], was used for this purpose. Briefly, ResNet was pre-trained for image recognition by using the ImageNet dataset [34] with over 1.2 million natural images sampling from 1,000 categories, yielding 75.3% top-1 test accuracy. The pretrained ResNet was able to predict the fMRI responses to videos with overall high and statistically significant accuracies throughout the visual cortex [27]. The ResNet consisted of 50 hidden layers of nonlinear computational units that encoded increasingly abstract and complex visual features. The first layer encoded location and orientation-selective visual features, whereas the last layer encoded semantic features that supported categorization. The layers in between encoded increasingly complex features through 16 residual blocks. Passing an image into ResNet yielded an activation value at each unit. Passing every frame of a movie into ResNet yielded an activation time series at each unit, indicating the time-varying representation of a specific feature in the movie. In this way, the feature representations of the training and testing movies could be extracted, as in previous studies [7, 27]. Here, we extracted the features from the first layer, the last layer, and the output layer for each of the 16 residual blocks in ResNet.

4.2.3 Feature dimension reduction

The feature space encoded in ResNet had a huge dimension over 10^6. This dimensionality could be reduced since individual features were not independent. For this purpose, principal component analysis (PCA) was applied first to each layer and then across layers. To define a set of principal components generalizable across various visual stimuli, a training movie as long as 12.54 hours was used to sample the original feature space. The corresponding feature representations were convolved with a canonical hemodynamic response function and then demeaned and divided by its standard deviation, yielding the standardized feature representation at each unit. Then, PCA was applied to the standardized feature representations from all units in each layer, as expressed as Eq. (1).

$$\boldsymbol{f}_l(\mathbf{x}) = \boldsymbol{f}_l^o(\mathbf{x})\mathbf{B}_l \qquad (1)$$

where $\boldsymbol{f}_l^o(\mathbf{x}) \in \mathbb{R}^{1\times p_l}$ stands for the standardized feature representation from layer l given a visual input $\mathbf{x}$, $\mathbf{B}_l \in \mathbb{R}^{p_l\times q_l}$ consists of the principal components (as unitary column vectors) for layer l, $\boldsymbol{f}_l(\mathbf{x}) \in \mathbb{R}^{1\times q_l}$ is the feature representation after reducing the dimension from p_l to q_l.

Due to the high dimensionality of the original feature space and the large number of video frames, we used an efficient singular value decomposition updating algorithm (or SVD-updating algorithm) [165, 166] to obtain the principal components $\mathbf{B}_l$. Briefly, the 12.54-hour training movie was divided into blocks, where each block was defined as an 8-min segment (i.e. a single fMRI session). The principal components of feature representations were first calculated for a block and then were incrementally updated with new blocks, by keeping >99% variance of the feature representations of every block.

Following the layer-wise dimension reduction, PCA was applied to the feature representations from all layers with SVD-updating algorithm, by keeping the principal components that explained >99% variance across layers for every block of visual stimuli. The final dimension reduction was implemented as Eq. (2).

$$\boldsymbol{f}(\mathbf{x}) = \boldsymbol{f}_{1:L}(\mathbf{x})\mathbf{B}_{1:L} \qquad (2)$$

where $\boldsymbol{f}_{1:L}(\mathbf{x}) = \left[\frac{\boldsymbol{f}_1(\mathbf{x})}{\sqrt{p_1}}, \ldots, \frac{\boldsymbol{f}_L(\mathbf{x})}{\sqrt{p_L}}\right]$ stands for the feature representations concatenated across L layers, $\mathbf{B}_{1:L}$ consists of the principal components of $\boldsymbol{f}_{1:L}(\mathbf{x})$ given the 12.54-hour training movie, and $\boldsymbol{f}(\mathbf{x}) \in \mathbb{R}^{1\times k}$ is the final dimension-reduced feature representation.

The principal components $\mathbf{B}_l$ and $\mathbf{B}_{1:L}$ together defined a dimension-reduced feature space, and their transpose defined the transformation to the original feature space. So, given any visual stimulus $\mathbf{x}$, its dimension-reduced feature representation could be obtained through Eqs. (1) and (2) with fixed $\mathbf{B}_l$ and $\mathbf{B}_{1:L}$. Once trained, the feature model including the feature dimension reduction, was assumed to be common to any subjects and any stimuli.

4.2.4 Voxel-wise linear response model

As the second part of the encoding model, a voxel-wise linear regression model was trained to predict the response $r_v(\mathbf{x})$ at voxel v evoked by the stimulus $\mathbf{x}$. In some previous studies [7, 8, 30], the encoding model for each voxel was based on a single layer in DNN that was relatively more predictive of the voxel's response than were other layers. Herein, we did not assume a one-to-one correspondence between a brain voxel and a ResNet layer. Instead, the feature representations from all layers were used (after dimension reduction) to predict each voxel's

response to video stimuli. After training, the regression coefficients of voxel-wise response models could still reveal the differential contributions of the features in different ResNet layers to each voxel [27, 167].

Mathematically, the linear response model was expressed by Eq. (3).

$$r_v(\mathbf{x}) = \boldsymbol{f}(\mathbf{x})\boldsymbol{w}_v + \varepsilon_v \qquad (3)$$

where $\boldsymbol{w}_v$ is a column vector of unknown regression coefficients specific to voxel v, and ε_v is the noise (unexplained by the model). Here, the noise was assumed to follow a Gaussian distribution with zero mean and variance equal to σ_v^2, i.e. $\varepsilon_v \sim \mathrm{N}(0, \sigma_v^2)$. Eq. (3) can be rewritten in vector/matrix notations as Eq. (4) for a finite set of visual stimuli (e.g. movie frames).

$$\boldsymbol{r}_v = \mathbf{F}\boldsymbol{w}_v + \boldsymbol{\varepsilon}_v \qquad (4)$$

where $\mathbf{F} \in \mathbb{R}^{n\times k}$ stands for the feature representations of n stimuli, $\boldsymbol{r}_v \in \mathbb{R}^{n\times 1}$ is the corresponding evoked responses, and $\boldsymbol{\varepsilon}_v \sim \mathrm{N}(0, \sigma_v^2\mathbf{I})$.

To estimate the regression coefficients $\boldsymbol{w}_v$ in Eq. (4), we used and compared two methods, both of which are subsequently described in a common framework of Bayesian inference. In the first method, we assumed the prior distribution of $\boldsymbol{w}_v$ as a zero-mean multivariate Gaussian distribution without using any knowledge from a model pretrained with previous data from the same or other subjects [151, 168]. With such a zero-mean prior, we maximized the posterior probability of $\boldsymbol{w}_v$ given the stimulus $\mathbf{x}$ and the fMRI response $r_v(\mathbf{x})$. In the second method, we assumed the prior distribution of $\boldsymbol{w}_v$ as a multivariate Gaussian distribution, whereas the mean was not zero but proportional to the regression coefficients in the pretrained model. As such, the prior was transferred from existing knowledge about the model as learned from existing data or other subjects (hereafter we referred to this prior as the transferred prior). The first method was used for training subject-specific encoding models with subject-specific training data. The second method was what we proposed for transferring encoding models across subjects, as illustrated in Fig. 4.1a.

4.2.5 Training the response model with the zero-mean prior

From Eq. (4), the likelihood of the response $\boldsymbol{r}_v$ given the unknown parameters $\boldsymbol{w}_v$ and the known feature representations $\mathbf{F}$ followed a multivariate Gaussian distribution, as Eq. (5).

$$p(\boldsymbol{r}_v|\boldsymbol{w}_v, \mathbf{F}) = \frac{1}{\sqrt{(2\pi\sigma_v^2)^n}} \exp\left\{-\frac{\|\boldsymbol{r}_v - \mathbf{F}\boldsymbol{w}_v\|_2^2}{2\sigma_v^2}\right\} \qquad (5)$$

In the framework of Bayesian inference, $\boldsymbol{w}_v$ was a multivariate random variable that followed a multivariate Gaussian distribution with a zero-mean, and an isotropic covariance $\boldsymbol{\Sigma}_v = s_v^2\mathbf{I}$, as expressed in Eq. (6).

$$p(\boldsymbol{w}_v) = \frac{1}{\sqrt{\left(2\pi s_v^2\right)^k}} \exp\left\{-\frac{\|\boldsymbol{w}_v\|_2^2}{2s_v^2}\right\} \qquad (6)$$

The prior distribution was independent of the visual input and thus its feature representations, i.e. $p(\boldsymbol{w}_v) = p(\boldsymbol{w}_v|\mathbf{F})$. Therefore, given $\mathbf{F}$ and $\boldsymbol{r}_v$, the posterior distribution of $\boldsymbol{w}_v$ was written as Eq. (7) according to the Bayes' rule.

$$p(\boldsymbol{w}_v|\boldsymbol{r}_v, \mathbf{F}) = \frac{p(\boldsymbol{r}_v|\boldsymbol{w}_v, \mathbf{F})p(w_v)}{p(\boldsymbol{r}_v|\mathbf{F})} \qquad (7)$$

where $p(\boldsymbol{r}_v|\mathbf{F})$ was constant since $\boldsymbol{r}_v$ and $\mathbf{F}$ were known. According to Eqs. (5), (6) and (7), the Bayesian estimation of $\boldsymbol{w}_v$ was obtained by maximizing the natural logarithm of its posterior probability, which was equivalent to minimizing the objective function as Eq. (8).

$$g(\boldsymbol{w}_v) = \frac{1}{n}\|\boldsymbol{r}_v - \mathbf{F}\boldsymbol{w}_v\|_2^2 + \lambda\|\boldsymbol{w}_v\|_2^2 \qquad (8)$$

where $\lambda = \frac{\sigma_v^2/n}{s_v^2}$. The analytical solution to minimizing (8) is as Eq. (9).

$$\widehat{\boldsymbol{w}}_v = (\mathbf{G} + \lambda\mathbf{I})^{-1}\,[\mathbf{F}]^{\mathrm{T}}\boldsymbol{r}_v/n \qquad (9)$$

where $\mathbf{G} = [\mathbf{F}]^{\mathrm{T}}\mathbf{F}/n$ is the covariance matrix of $\mathbf{F}$.

4.2.6 Training the response model with the transferred prior

If a pretrained model, $\boldsymbol{w}_v^0$, was available, we could use this model to derive more informative and precise prior knowledge about $\boldsymbol{w}_v$. Specifically, $\boldsymbol{w}_v$ was assumed to follow a multivariate Gaussian distribution, of which the mean was $\alpha\boldsymbol{w}_v^0$ (α is a non-negative factor) and the covariance was $\boldsymbol{\Sigma}_v = s_v^2\mathbf{I}$. The prior probability of $\boldsymbol{w}_v$ was as Eq. (10).

$$p(\boldsymbol{w}_v) = \frac{1}{\sqrt{\left(2\pi s_v^2\right)^k}} \exp\left\{-\frac{\|\boldsymbol{w}_v - \alpha\boldsymbol{w}_v^0\|_2^2}{2s_v^2}\right\} \qquad (10)$$

Here, the prior was transferred from a pretrained model (namely the transferred prior), and was used to constrain the mean of the model to be trained with new data and/or for a new subject. According to Eqs. (5), (7) and (10), maximizing the posterior probability of $\boldsymbol{w}_v$ was equivalent to minimizing the following objective function.

$$g(\boldsymbol{w}_v) = \frac{1}{n}\|\boldsymbol{r}_v - \mathbf{F}\boldsymbol{w}_v\|_2^2 + \lambda\|\boldsymbol{w}_v - \alpha\boldsymbol{w}_v^0\|_2^2 \qquad (11)$$

where $\lambda = \frac{\sigma_v^2/n}{s_v^2}$. Note that if $\alpha = 0$, this objective function becomes equivalent to Eq. (8). The objective function could be reformatted as Eq. (12), where $a = \alpha\lambda$, $b = (1-\alpha)\lambda$, and c is a constant.

$$g(\boldsymbol{w}_v) = \frac{1}{n}\|\boldsymbol{r}_v - \mathbf{F}\boldsymbol{w}_v\|_2^2 + a\|\boldsymbol{w}_v - \boldsymbol{w}_v^0\|_2^2 + b\|\boldsymbol{w}_v\|_2^2 + c \qquad (12)$$

In this function, the first term stands for the mean square error of model fitting, the second term stands for the deviation from the prior model, $\boldsymbol{w}_v^0$, and the third term had a similar regularization effect as that in Eq. (8). The analytical solution to minimizing (12) was as Eq. (13).

$$\widehat{\boldsymbol{w}}_v = [\mathbf{G} + (a+b)\mathbf{I}]^{-1}(a\boldsymbol{w}_v^0 + [\mathbf{F}]^{\mathrm{T}}\boldsymbol{r}_v/n) \qquad (13)$$

where $\mathbf{G} = [\mathbf{F}]^{\mathrm{T}}\mathbf{F}/n$ is the covariance matrix of $\mathbf{F}$.

4.2.7 Choosing hyper-parameters with cross-validation

The hyper-parameters λ in Eq. (9) or (a, b) in Eq. (13) were determined for each voxel by four-fold cross-validation [169]. Specifically, the training video-fMRI dataset was divided into four subsets of equal size: three for the model estimation, and one for the model validation. The validation accuracy was measured as the correlation between the predicted and measured cortical responses. The validation was repeated four times such that each subset was used once for validation. The validation accuracy was averaged across the four repeats. Finally, the hyper-parameters were chosen such that the average validation accuracy was maximal.

4.2.8 Testing the encoding performance with the testing movie

Once voxel-wise encoding models were trained, we evaluated the accuracy of using the trained models to predict the cortical responses to the testing movies, which were not used for training the encoding models. The prediction accuracy was quantified as the correlation (r) between the predicted and observed fMRI responses at each voxel given the testing movie. Since the testing movie included five different 8-min sessions with entirely different content, the prediction accuracy was evaluated separately for each session and then averaged across sessions. The significance of the average voxel-wise prediction accuracy was evaluated with a block-permutation test [59] with a block length of 30 seconds (corrected at false discovery rate (FDR) $q < 0.01$), as used in our prior study [7, 27].

4.2.9 Evaluating the encoding models without any transferred prior

For a specific subject, when the voxel-wise encoding model was estimated without any prior information from existing models pre-trained for other subjects, the estimated model was entirely based on the subject-specific training data. In this case, we evaluated how the encoding performance depended on the size of the training data.

To do so, we trained the encoding models for Subject JY using a varying part of the 10.4-hour training data. The data used for model training ranged from 16 minutes to 10.4 hours. For such models trained with varying lengths of data, we tested their individual performance in predicting the responses to the 40-min testing movie. We calculated the percentage of predictable voxels (i.e. significant with the block-permutation test) out of the total number of cortical voxels, and evaluated it as a function of the size of the training data. We also evaluated the histogram of the prediction accuracy for all predictable voxels, and calculated the overall prediction accuracy in regions of interest (ROIs) [62] by averaging across voxels within ROIs.

4.2.10 Evaluating the encoding models with the transferred prior

When the voxel-wise encoding model was trained with the prior transferred from a pretrained model, the parameters in the new model depended on both the pretrained model and the new training data. As such, one might not require so many training data to train the model as required without the transferred prior.

We used this strategy for transferring encoding models from one subject to another. Specifically, we trained the models from scratch based on the 10.4-hour training data from one subject (JY), and used the trained models as the model prior for other subjects (XF and XL). With this prior model from Subject JY, we trained the encoding models for Subject XF and XL based on either short (16 minutes, i.e. two 8-min sessions) or long (2.13 hours, i.e. 16 sessions) training data specific to them. Note that the movie used for training the prior model in Subject JY was different from either the training or testing movies for Subject XL and XF. With either short or long training data, we evaluated the encoding performance in predicting the responses to the testing movie for Subject XF and XL. For comparison, we also evaluated the encoding models trained with the same training data from Subject XF and XL without using any transferred knowledge from Subject JY, or the prior models from Subject JY without being retrained with any data from Subject XF and XL. The comparison was made with respect to the number of predictable voxels

and the voxel-wise prediction accuracy (after converting the correlation coefficients to the z scores with the Fisher's r-to-z transform). The model comparison was conducted repeatedly when the models under comparison were trained (or tested) with distinct parts of the training (or testing) movie. Between different models, their difference in encoding performance was tested for significance by applying one-sample t-test to the repeatedly measured prediction accuracy (corrected at false discovery rate (FDR) q<0.01).

We also conducted similar analyses by using Subject JY as the target subject, for whom the encoding models were trained with prior knowledge transferred from the encoding models trained with data from Subject XL or XF. Note that the prior models were trained with 1.87-hour training data, and then were refined with 16min data from the target subject. Note that the movie used for training the prior model was different from the movie for refining the prior model for the target subject.

4.2.11 Hyperalignment between subjects

We also explored whether transferring encoding models from one subject to another would also benefit from performing functional hyperalignment as an additional preprocessing step. Specifically, we used the searchlight hyperalignment algorithm [164] to correct for the individual difference in the fine-scale functional representation beyond what could be accounted for by anatomical alignment [49]. Given the 16-min alignment movie, the fMRI responses within a searchlight (with a radius of 20mm) were viewed as a high-dimensional vector that varied in time. A Procrustes transformation [170] was optimized to align high-dimensional response patterns from one subject to another [164].

To evaluate the effect of hyperalignment in transferring encoding models across subjects, we performed the searchlight hyperalignment from Subject JY to Subject XL and XF. Then we applied the functional hyperalignment to the encoding models trained for the source subject (Subject JY) to give rise to the prior models that were used for training the encoding models for the target subject (Subject XL or XF). The encoding performance of the resulting models was evaluated and compared with those without hyperalignment. The difference in the encoding performance was addressed with respect to the number of predictable voxels and the voxel-wise prediction accuracy, and was tested for significance with one-sample t-test corrected at false discovery rate (FDR) q<0.01.

4.2.12 Training group-level encoding models with online learning

Here, we describe an online learning algorithm [163] to train group-level encoding models based on different video-fMRI data acquired from different subjects, by extending the concept of online implementation for the Levenberg-Marquardt algorithm [171]. The central idea is to update the encoding models trained with existing data based on the data that become newly available, as illustrated in Fig. 4.1b.

Suppose that existing training data are available for a set of visual stimuli, $\mathbf{X}^0$ (n^0 samples). Let $\mathbf{F}^0$ be the corresponding feature representations after dimension reduction, $\boldsymbol{r}_v^0$ be the responses at voxel v. Let $\boldsymbol{w}_v^0$ be the regression parameters in the voxel-specific encoding models trained with $\mathbf{F}^0$ and $\boldsymbol{r}_v^0$ according to Eq. (9). Given incremental training data, $\mathbf{X}^1$ (n^1 samples), $\mathbf{F}^1$ and $\boldsymbol{r}_v^1$, the parameters in the updated encoding model can be obtained by minimizing the objective function below.

$$g_{\mathrm{G}}(\boldsymbol{w}_v) = \frac{1}{n^0+n^1}\left\|\begin{bmatrix}\boldsymbol{r}_v^0\\ \boldsymbol{r}_v^1\end{bmatrix} - \begin{bmatrix}\mathbf{F}^0\\ \mathbf{F}^1\end{bmatrix}\boldsymbol{w}_v\right\|_2^2 + \lambda\|\boldsymbol{w}_v\|_2^2 \qquad (14)$$

The optimal solution is expressed as Eq. (15).

$$\boldsymbol{w}_v = (1-\theta)(\mathbf{G}+\lambda\mathbf{I})^{-1}(\mathbf{G}^0+\lambda^0\mathbf{I})\boldsymbol{w}_v^0 + \theta(\mathbf{G}+\lambda\mathbf{I})^{-1}\,[\mathbf{F}^1]^{\mathrm{T}}\boldsymbol{r}_v^1/n^1 \qquad (15)$$

where $\mathbf{G}^0 = [\mathbf{F}^0]^{\mathrm{T}}\mathbf{F}^0/n^0$ and $\mathbf{G}^1 = [\mathbf{F}^1]^{\mathrm{T}}\mathbf{F}^1/n^1$ are the covariance matrices of $\mathbf{F}^0$ and $\mathbf{F}^1$, respectively; $\mathbf{G} = (1-\theta)\mathbf{G}^0 + \theta\mathbf{G}^1$ is their weighted sum where the parameter θ specifies the relative weighting of the new data and the previous data. See [28] for the derivation of Eq. (15). In this study, θ was set as the ratio of the corresponding sample sizes, i.e. $\theta = \frac{n^1}{n^0+n^1}$. As such, the samples in the new data were assumed to be as important as those in the previous data.

According to Eq. (15), the encoding model could be incrementally updated by incorporating new data without training the model from scratch. See **Algorithm 1** Table 4.1 for the updating rules. As more and more data was used for model training, the encoding model was expected to converge, as $(\mathbf{G}+\lambda\mathbf{I})^{-1}(\mathbf{G}^0+\lambda^0\mathbf{I}) \longrightarrow \mathbf{I}$ and $\theta \longrightarrow 0$. When it was used to utilize the growing training data from different subjects, this algorithm converged to the group-level encoding models.

As a proof of concept, we trained group-level encoding models by incrementally updating the models with 16-min video-fMRI training data sampled from each of the three subjects in the group. Before each update, the incremental fMRI data was functionally aligned to the data already used

to train the existing models. After the encoding models were trained with all the training data combined across all the subjects, we evaluated their prediction performance given the testing movie for each subject. The prediction accuracy of the group-level encoding models was averaged across subjects. We then compared the prediction performance before and after every update by incorporating new data.

4.3 Results

In recent studies, DNNs driven by image or action recognition were shown to be able to model and predict cortical responses to natural picture or video stimuli [7-10, 22, 27, 30, 155]. This ability rested upon encoding models, in which non-linear features were extracted from visual stimuli through DNNs and the extracted features were projected onto stimulus-evoked responses at individual locations through linear regression. Herein, we investigated the amount of data needed to train DNN-based encoding models in individual subjects, and developed new methods for transferring and generalizing encoding models across subjects without requiring extensive data from single subjects.

4.3.1 Encoding performance depended on the size of the training data

In this study, we focused on a specific DNN (i.e. ResNet) – a feed-forward convolutional neural network (CNN) pre-trained for image recognition [21]. The ResNet included 50 successive layers of computational units, extracting around 10^6 non-linear visual features. This huge dimensionality could be reduced by two orders of magnitude, by applying PCA first to every layer and then across all layers. The reduced feature representations were able to capture 99% of the variance of the original features in every layer.

Despite the reduction of the feature dimensionality, training a linear regression model to project the feature representations onto the fMRI response at each voxel still required a large amount of data if the model was estimated solely based on the training data without any informative prior knowledge. For such encoding models, we evaluated the effects of the size of the training data on the models' encoding performance in terms of the accuracy of predicting the responses to the testing movie, of which the data were not used for training to ensure unbiased testing. When trained with 10.4 hours of video-fMRI data, the prediction accuracy of the encoding models was statistically significant (permutation test, FDR $q<0.01$) for nearly the entire visual

cortex (Fig. 4.2.a). The number of predictable voxels and the prediction accuracy were notably reduced as the training data were reduced to 5.87 hours, 2.13 hours, or 16 minutes (Fig. 4.2.b). With increasing sizes of training data, the predictable areas increased monotonically, from about 20% (with 16-min of training data) to >40% (with 10.4-hour of data) of the cortical surface (Fig. 4.2.c). The average prediction accuracies, although varying across regions of interest (ROIs), showed an increasing trend as a growing amount of data were used for model training (Fig. 4.2.d). It appeared that the trend did not stop at 10.4 hours, suggesting a sub-optimal encoding model even if trained with such a large set of training data. Therefore, training encoding models for a single subject purely relying on training data would require at least 10 hours of video-fMRI data from the same subject.

4.3.2 Transferring encoding models across subjects through Bayesian inference

To mitigate this need for large training data from every subject, we asked whether the encoding models already trained with a large amount of training data could be utilized as the prior information for training the encoding models in a new subject with much less training data. To address this question, we used the encoding models trained with 10.4 hours of training data from Subject JY as *a priori* models for Subject XF and XL. A Bayesian inference method was used to utilize such prior models for training the encoding models for Subject XF and XL with either 16-min or 2.13-hour training data from these two subjects. The resulting encoding models were compared with those trained without using any prior models with the same amount of training data in terms of their accuracies in predicting the responses to the testing movie.

Fig. 4.3 shows the results for the model comparison in Subject XF. When the training data were as limited as 16 minutes, the encoding models trained with the prior modeled transferred from another significantly outperformed those without using the prior (Fig. 4.3.a). With the prior model, the predictable cortical areas were 26% of the entire cortex, nearly twice as large as the predictable areas without the prior (14.9% of the entire cortex). Within the predictable areas, the prediction accuracy was also significantly higher with the prior model ($\Delta z = 0.155 \pm 0.0006$, one-sample t-test, $p<10^{-5}$) (Fig. 4.3.a). The difference in voxel-wise prediction accuracy was significant (one-sample t-test, $p<0.01$) in most of the visual areas, especially for those in the ventral stream (Fig. 4.3.a). The advantage of using the prior model largely diminished when 2.13-hour training data were used for training the encoding models (Fig. 4.3.b). Although larger training data

improved the model performance, the improvement was much more notable for the method when the prior model was not utilized. In that case, the predictable area increased from 14.9% to 26.7% of the cortex ($p=6.5\times10^{-5}$, paired t-test). When the prior model was utilized, the predictable area increased from 26.0% to 28.5% ($p=0.017$, paired t-test), and the prediction accuracy only improved marginally (Fig. 4.3.b). Similar results were observed when transferring from Subject JY to Subject XL [28], as well as across other pairs of subjects [28]. It was noteworthy that the prediction accuracy of the transferred encoding model with 16-min fMRI data was comparable to the non-transferred models with 2.13-hour fMRI data (Fig. 4.3).

We also asked whether the better performance of the encoding models with the transferred prior was entirely attributable to the prior models from a different subject, or it could be in part attributable to the information in the training data specific to the target subject. To address this question, we directly used the prior models (trained with data from Subject JY) to predict the cortical responses to the testing movie in Subject XL and XF. Even without any further training, the prior models themselves yielded high prediction accuracy for widespread cortical areas in Subject XF for whom the models were not trained (Fig. 4.4.a). When the prior models were fine-tuned with a limited amount (16-min) of training data specific to the target subject, the encoding performance was further improved (Fig. 4.4.b). The improvement was greater when more (2.13-hour) training data were utilized for refining the encoding models (Fig. 4.4.c). Similar results were also observed in another subject [28]. Hence, Bayesian inference to transfer encoding models across subjects could help train the encoding models for new subjects without requiring extensive training data from them. The subject-specific training data served to tailor the encoding models from the source subject towards the target subject.

4.3.3 Functional alignment better accounted for individual differences

Transferring encoding models across subjects were based on the assumption that the models and data from individual subjects were co-registered. Typically, the co-registration was based on anatomical features (i.e. anatomical alignment) [49]. We expected that searchlight hyperalignment of multi-voxel responses could better co-register the fine-grained representational space on the cortical surface [164] to improve the efficacy of transferring the encoding models across subjects.

Therefore, we performed searchlight hyperalignment such that Subject JY's fMRI responses to the alignment movie were aligned to the other subjects' responses to the same movie. After applying the same alignment to the encoding models trained for Subject JY, we used the aligned encoding models as the prior model for training the encoding models for Subject XF or XL with 16-min training datasets from each of them. It turned out that using the functional alignment as a preprocessing step further improved the performance of the transferred encoding models. For Subject XF, the model-predictable areas increased from 26% to 27.8% ($p=9.7\times10^{-4}$, paired t-test), and the prediction accuracy also increased, especially for the extrastriate visual areas (Fig. 4.5).

4.3.4 Group-level encoding models

We further explored and tested an online learning strategy to train the encoding models for a group of subjects by incrementally using data from different subjects for model training. Basically, incremental neural data (16 minutes) was obtained from a new subject with new visual stimuli, and was used to update the existing encoding models (Fig. 4.6a). Such learning strategy allowed training group-level encoding models. The models significantly predicted the cortical response to novel testing movie for each subject (Fig. 4.6b). With every incremental update, the encoding models predicted wider cortical areas that increasingly covered 18.4%, 21.72%, and 24.27% of the cortex, and achieved higher prediction accuracies within the predictable areas (first update: $\Delta z = 0.05 \pm 0.0006$, $p<10^{-5}$; second update: $\Delta z = 0.036 \pm 0.00034$, $p<10^{-5}$, one-sample t-test) (Fig. 4.6.b). Meanwhile, the group-level encoding models exhibited similar predictability across individual subjects [28].

4.4 Discussion

Methods and Materials In this article, we have described methods to transfer and generalize encoding models of natural vision across human subjects. Central to our methods is the idea of taking the models learnt from data from one subject (or a group of subjects) as the prior models for training the models for a new subject (or a new group of subjects). This idea, implemented in the framework of Bayesian inference, allows to train subject-specific encoding models with a much less amount of training data than otherwise required if training the models from scratch without considering any pretrained model prior. The efficacy of this method, as demonstrated in

this paper, suggests that different subjects share largely similar cortical representations of vision [51, 100, 156, 172]. It has also led us to develop a method to train encoding models generalizable for a population by incrementally learning from different training data collected from different subjects.

The methods are described in the context of using DNN as a feature model, but they are also valuable and applicable to other models of visual or conceptual features [24-26, 60]. In general, the larger the feature space is, the more data is required for training the model that relates the features to brain responses in natural vision. DNNs attempt to extract hierarchical visual features in many levels of complexity, and thus it is so-far most data demanding to model their relationships to the visual cortex. Nevertheless, DNNs are of increasing interest for natural vision [1-3]. Recent studies have shown that DNNs, especially convolutional neural networks for image recognition [19-21], preserve the representational geometry in object-sensitive visual areas [9, 10, 22], and predicts neural and fMRI responses to natural picture or video stimuli [7, 8, 30, 155], suggesting their close relevance to how the brain organizes and processes visual information. DNNs also open new opportunities for mapping the visual cortex, including the cortical hierarchy of spatial and temporal processing [7, 8, 22, 30], category representation and organization [10, 27], visual-field maps [7, 30], all by using a single experimental paradigm with natural visual stimuli. It is even possible to use DNNs for decoding visual perception or imagery [7, 56]. Such mapping, encoding, and decoding capabilities all require a large amount of data from single subjects in order to train subject-specific models. Results in this study suggest that even 10 hours of fMRI data in response to diverse movie stimuli may still be insufficient for DNN-based encoding models (Fig. 4.2). Therefore, it is difficult to generalize the models established with data from few subjects to a large number of subjects or patients for a variety of potential applications.

The methods developed in this study fill this gap, allowing DNN-based encoding models to be trained for individual subjects without the need to collect substantial training data from them. As long as models have been already trained with a large amount of data from existing subjects or previous studies, such models can be utilized as the prior models for a new subject and be updated with additional data from this subject. Results in this study demonstrate that with prior models, encoding models can be trained with 16-min video-fMRI data from a single subject to reach comparable encoding performance as the models otherwise trained with over two hours of data but without utilizing any prior models (Fig. 4.3). Apparently, data acquisition for 16 minutes

readily fit into the time constraint of most fMRI studies. With the method described herein, it is thus realistic to train encoding models to effectively map and characterize visual representations in many subjects or patients for basic or clinical neuroscience research. The future application to patients with various cortical visual impairments, e.g. facial aphasia, has the potential to provide new insights to such diseases and their progression.

The methods developed for transferring encoding models across subjects might also be usable to transfer such models across imaging studies with different spatial resolution. The fMRI data in this study are of relatively low resolution (3.5mm). Higher resolution about 1mm is readily attainable with fMRI in higher field strengths (e.g. 7T or above) [173]. Functional images in different resolution reflect representations in different spatial scales. High-field and high-resolution fMRI that resolves representations in the level of cortical columns or layers is of particular interest [173, 174]; but prolonged fMRI scans in high-field face challenges, e.g. head motion and susceptibility artifacts as well as safety concern of RF power deposition. Transferring encoding models trained with 3-T fMRI data in lower resolution to 7-T fMRI data in higher resolution potentially enables higher throughput with limited datasets. Note that transferring the encoding models is not simply duplicating the models across subjects or studies. Instead, new data acquired from different subjects or with different resolution serve to reshape the prior models to fit the new information in specific subjects or representational scales. It is perhaps even conceivable to use the method in this study to transfer encoding models trained with fMRI data to those with neurophysiological responses observable with recordings of unit activity, local field potentials, and electrocorticograms. As such, it has the potential to compare and converge neural coding in different spatial and temporal scales. However, such a potential is speculative and awaits verification in future studies.

This study also supports an extendable strategy for training population-wide encoding models by collecting data from a large group of subjects. In most of the current imaging studies, different subjects undergo the same stimuli or tasks with the same experiment paradigm and the same acquisition protocol [175]. Such study design allows for more convenient group-level statistics, more generalizable findings, and easier comparison across individuals. However, if one has to collect substantial data from each subject, it is practical too expensive or unrealistic to do so for a large number of subjects. An alternative strategy is to design a study for a large number of subjects, but only collect imaging data from subjects undergoing different visual stimuli, e.g.

watching different videos. For the population as a whole, data with a large and diverse set of stimuli become available. The methods described herein lay the technical foundation to combine the data across subjects for training population-wide encoding models. This strategy may be further complemented by also using a small set of stimuli (e.g. 16-min video stimuli) common for all subjects. Such stimuli can be used to functionally align the data from different subjects to account for individual differences (Fig. 4.6) [100, 156, 164]. It also provides comparable testing data to assess individual differences.

In addition, our methods allow population-wide encoding models to be trained incrementally. For a study that involves data acquisition from many subjects, data are larger and growing. It is perhaps an unfavorable strategy to analyze the population data only after data are available from all subjects. Not only is it inefficient, analyzing the population data as a whole requires substantial computing resources – a common challenge for "big data". Using online learning [163], the methods described herein allows models to be trained and refined as data acquisition progresses. Researchers can examine the evolution of the trained models, and decide whether the models have converged to avoid further data acquisition. As population-wide encoding models become available, it is more desirable to use them as the prior models for training encoding models for specific subjects, or another population. Population-wide models are expected to be more generalizable than models trained from one or few subjects, making the prior models more valid and applicable for a wide group of subjects or patients.

Beyond the methods described in this paper, the notion of transferring encoding models across subjects may be substantiated with further methodological development. In this study, the encoding parameters in the prior model was used to constrain the mean of the parameters in a new model, whereas the covariance of the parameters were assumed to be isotropic. As such, all the parameters were assumed to bear different means but the same variance while being independent of each other. The assumption of independence was valid, because the feature space was reduced to a lower dimension, and was represented by its (orthogonal) principal components. The assumption of isotropic variance might be replaced by a more general covariance structure, in which the prior variance is allowed to be different for the parameters of individual features. Although it is possible to estimate the prior variance from the data, it requires a larger amount of training data and iterative optimization to estimate both the model parameters and their prior (anisotropic) variances for the maximum posterior probability [176]. The demand for data and

computation is what we aim to mitigate. Therefore, our assumption of isotropic variance is a legitimate choice, even though it may or may not be optimal.

In this study, we also assume a voxel-wise correspondence between one brain and another [51]. This assumption may not be optimal given the individual differences in the brain's structure and function [100, 156]. In addition to anatomical alignment [49], functional hyperalignment [164] is helpful to partly account for the individual differences, before transferring voxel-wise encoding models across subjects. It is also likely helpful to statistically summarize the prior model across neighboring voxels, or in a region that contains the target voxel. Refinement of the algorithms for transferring encoding models awaits future studies.

Lastly, this study focuses exclusively on natural vision. However, the methods developed are anticipated to serve well for more general purposes, including natural language processing, speech and hearing.

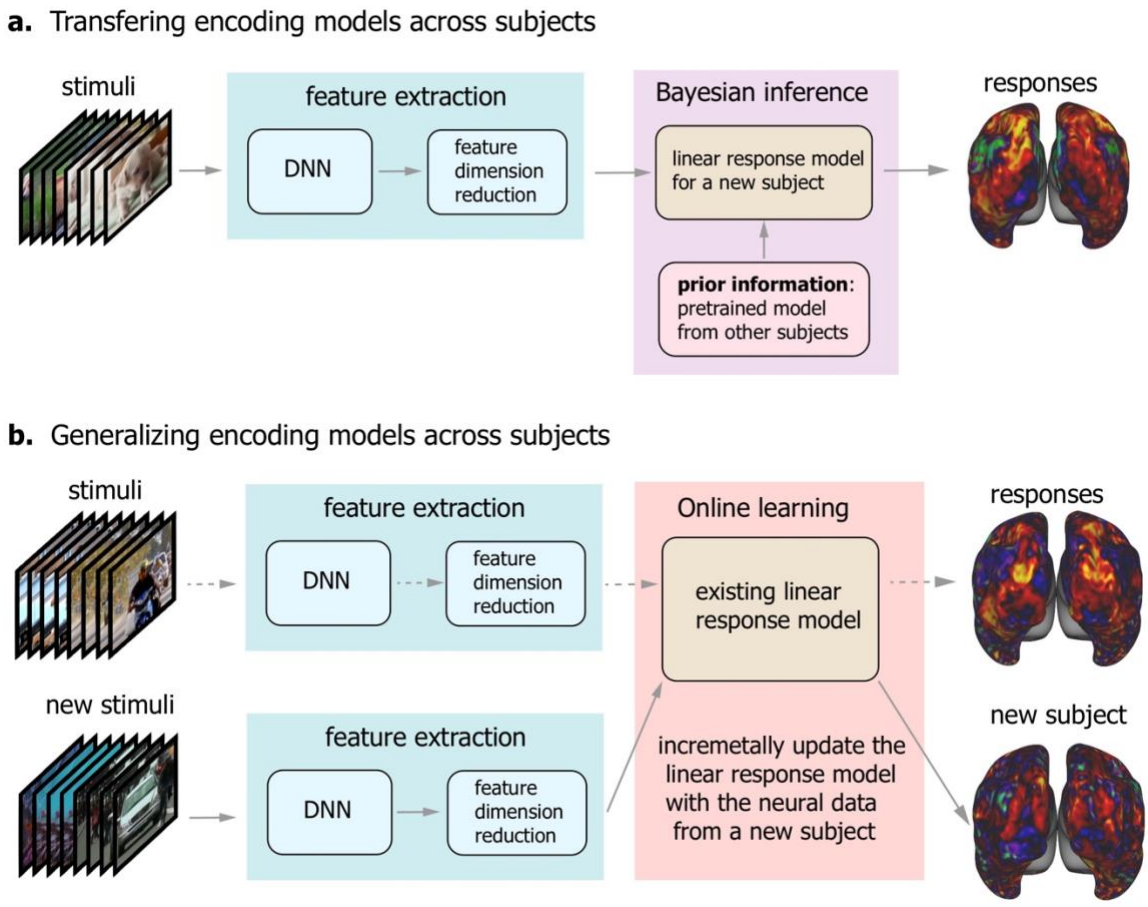

Figure 4.1 **Schemes of transferring and generalizing DNN-based neural encoding models across subjects.** (a) Transferring encoding models across subjects. The encoding model comprises the nonlinear feature model and the linear response model. In the feature model, the feature representation is extract from the visual stimuli through the deep neural network (DNN), and followed by the feature dimension reduction. In the response model, the model parameters are estimated by using Bayesian inference with subject-specific neural data as well as a prior model trained from other subjects. (b) Generalizing encoding models across subjects. The dash arrows indicate the existing encoding model trained with the data from a group of subjects. The existing model can be incrementally updated by using the new data from a new subject with an online learning algorithm. In the scheme, the feature model is common any subjects and any stimuli, and the response model will be updated when new subject data is available.

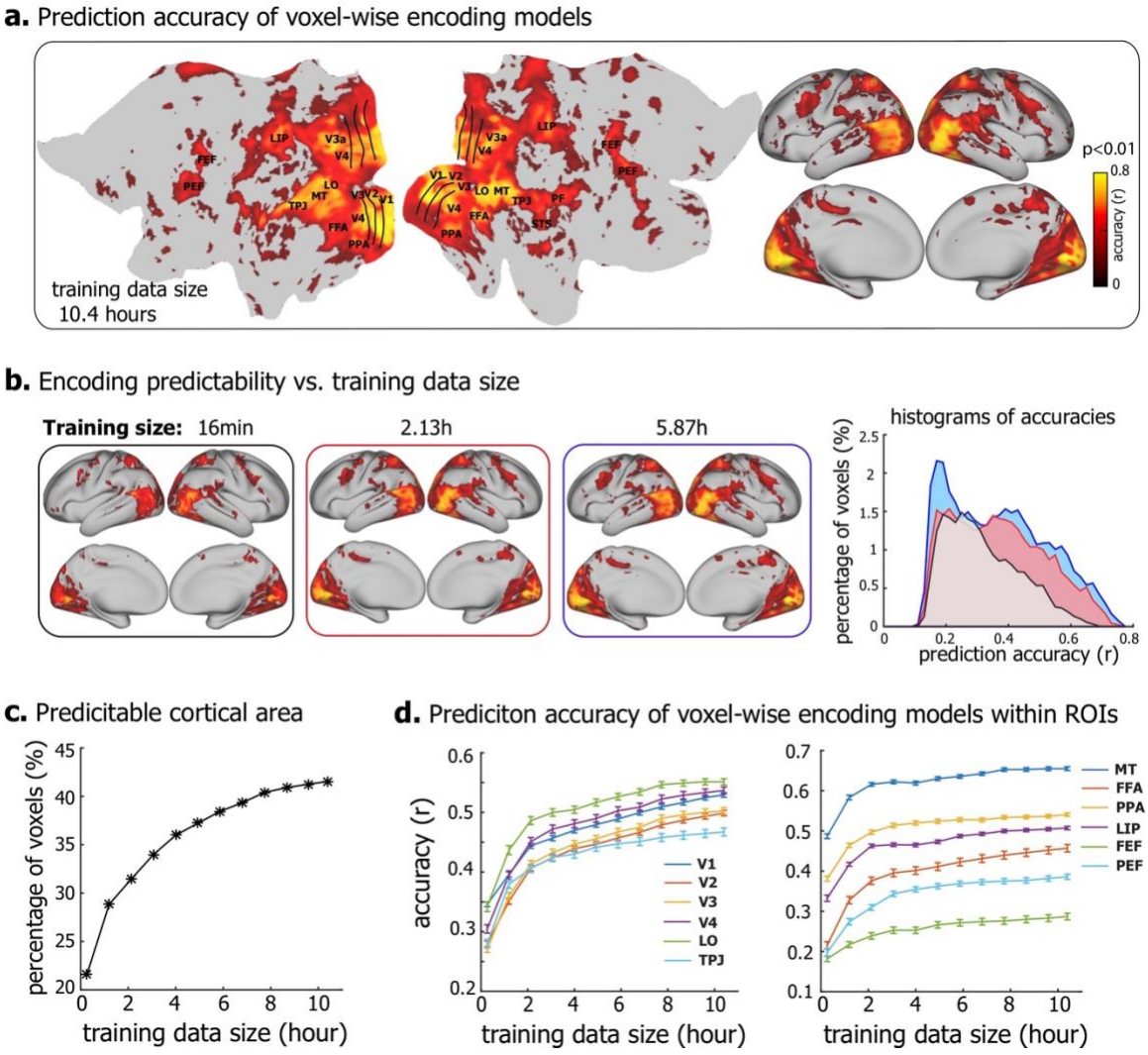

Figure 4.2 **DNN-based neural encoding models for Subject JY.** (a) Performance of neural encoding models (trained with 10.4-hour data) in predicting the cortical responses to novel testing movies. The accuracy is measured by the average Pearson's correlation coefficient (r) between the predicted and the observed fMRI responses across five testing movies (permutation test, q<0.01 after correction for multiple testing using the false discovery rate (FDR) method). The prediction accuracy is visualized on both flat (left) and inflated (right) cortical surfaces. (b) Prediction accuracy of encoding models trained with less training data, i.e. 16min, 2.13h, and 5.87h. The right is the histograms of prediction accuracies. The x-axis is the prediction accuracy ranging from 0 to 0.8, divided into bins of length $\Delta r = 0.02$, the y-axis is the percentages of predictable voxels in the cortex within accuracy bins. (c) The percentage of predictable voxels as a function of training data size ranging from 16min to 10.4 hours. (d) ROI-level prediction accuracies as functions of the training data size. The error bar indicates the standard error across voxels.

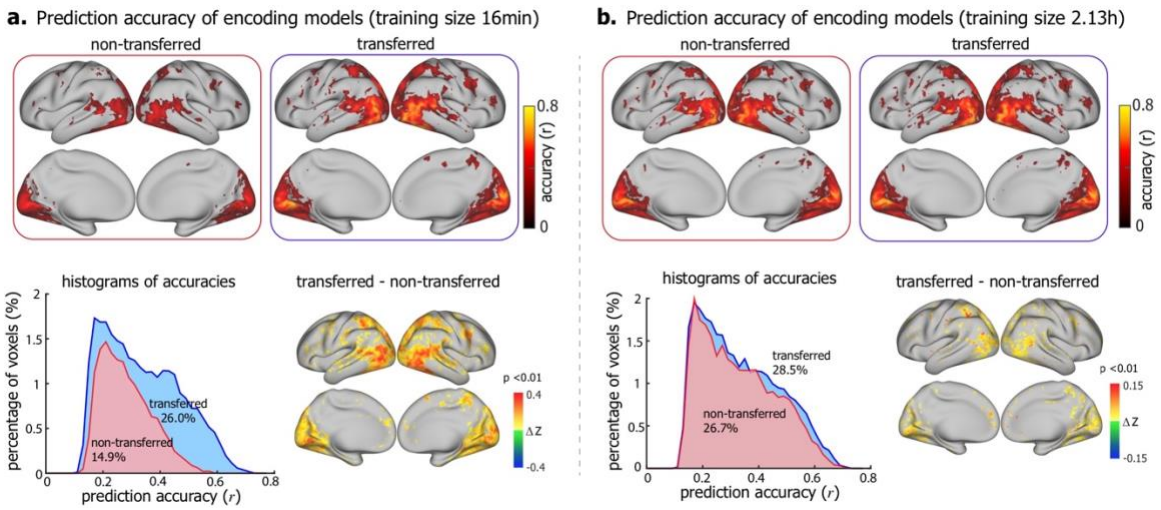

Figure 4.3 **Comparison between the encoding models that utilized the prior models transferred from a different subject (transferred) versus those without using any transferred prior (non-transferred).** Voxel-wise prediction accuracy of encoding models trained with 16min (a) and 2.13h (b) video-fMRI data (permutation test, corrected at FDR q<0.01). The top shows the voxel-wise prediction accuracy of the encoding models with the prior transferred from a pretrained model (right) and the encoding models without any transferred prior (left). The bottom left is the histograms of their respective prediction accuracies. The numbers are the total percentages of predictable voxels. The bottom right is the difference of prediction accuracy (Fisher's z-transformation of r, i.e. $z = \operatorname{arctanh}(r)$) between the encoding models with the transferred prior and those without any transferred prior (one-sample t-test, p<0.01). The figure shows the results for transferring from Subject JY to Subject XF, see Supplementary Figure S1 and S2 in [28] for other subjects.

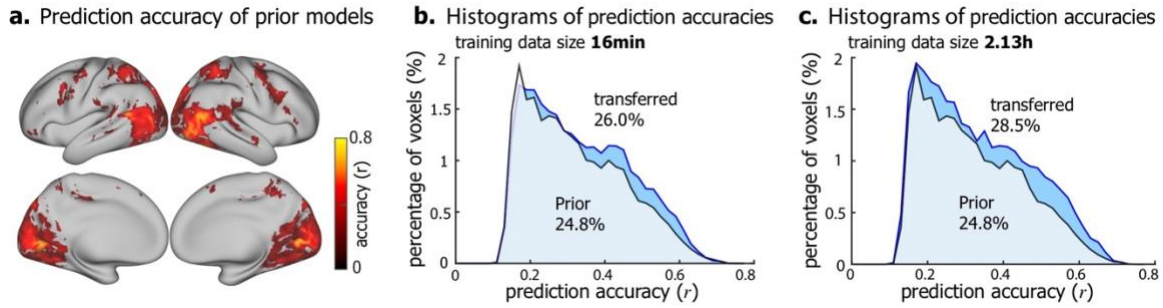

Figure 4.4 **Comparison between the encoding models that were refined from the prior models transferred from a different subject (transferred) versus the prior encoding models (prior).** (a) Voxel-wise prediction accuracy by directly using the prior encoding models (from Subject JY) to predict the responses to novel testing movies for Subject XF (permutation test, corrected at FDR $q<0.01$). (b) and (c) show the histograms of prediction accuracies of the encoding models that were transferred from the prior encoding models (blue) and the prior encoding models [149] trained with 16min (b) and 2.13h (c) training data, respectively. See Supplementary Figure S4 in [28] for Subject XL.

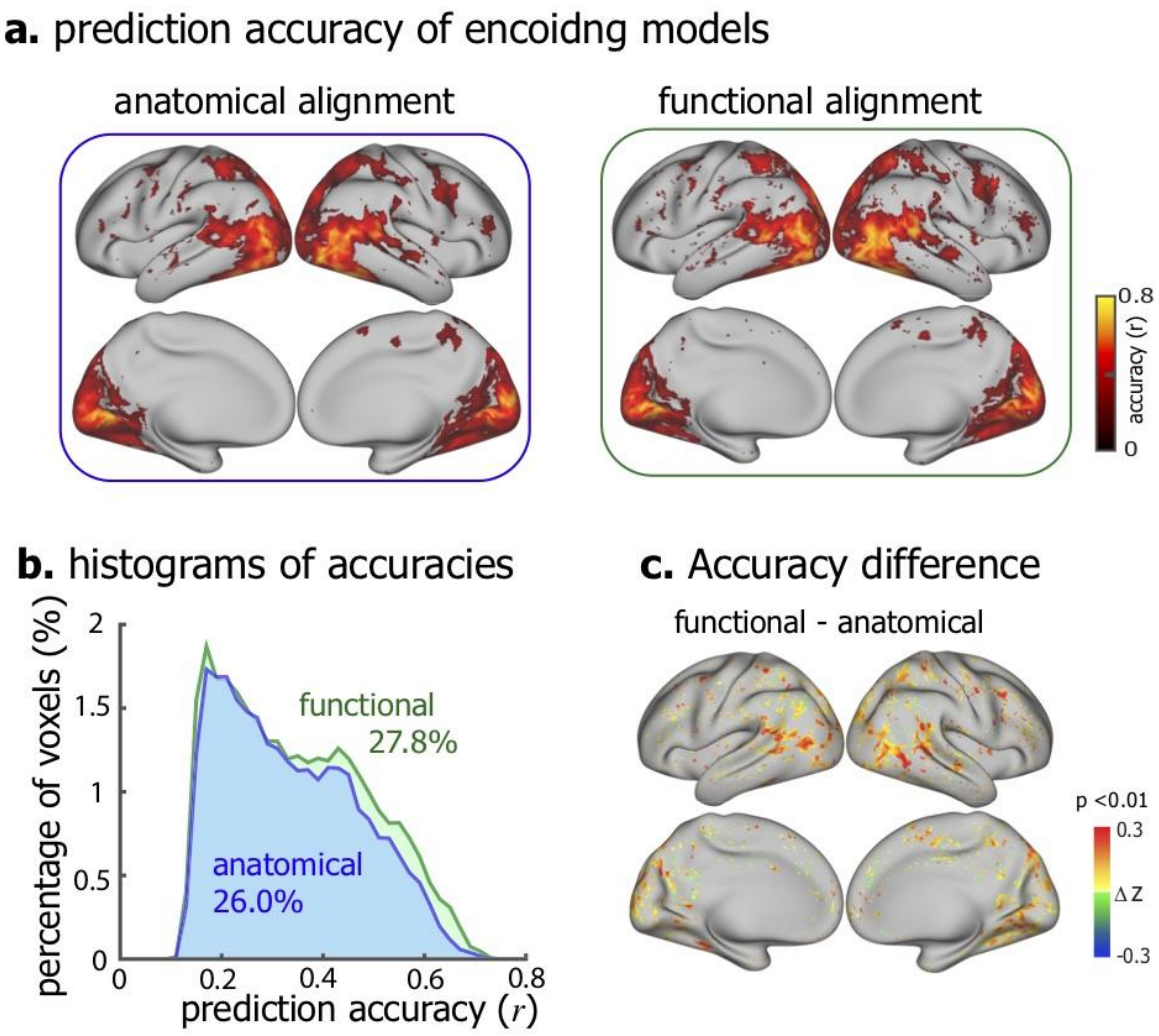

Figure 4.5 **Comparison of the encoding models that were transferred from prior models with anatomical versus functional alignment**. (a) Voxel-wise prediction accuracy of the encoding models based on anatomical alignment (left) and functional alignment (right) (permutation test, corrected at FDR q<0.01). (b) The histograms of prediction accuracies of anatomically aligned (blue) and functionally aligned [149] transferred encoding models. The colored numbers are the total percentages of predictable voxels. (c) The voxel-wise difference in prediction accuracy (Fisher's z-transformation of r, i.e. $z = \operatorname{arctanh}(r)$) between functional alignment and anatomical alignment (one-sample t-test, p<0.01). The figure shows the results for Subject XF, see Supplementary Figure S5 in [28] for Subject XL.

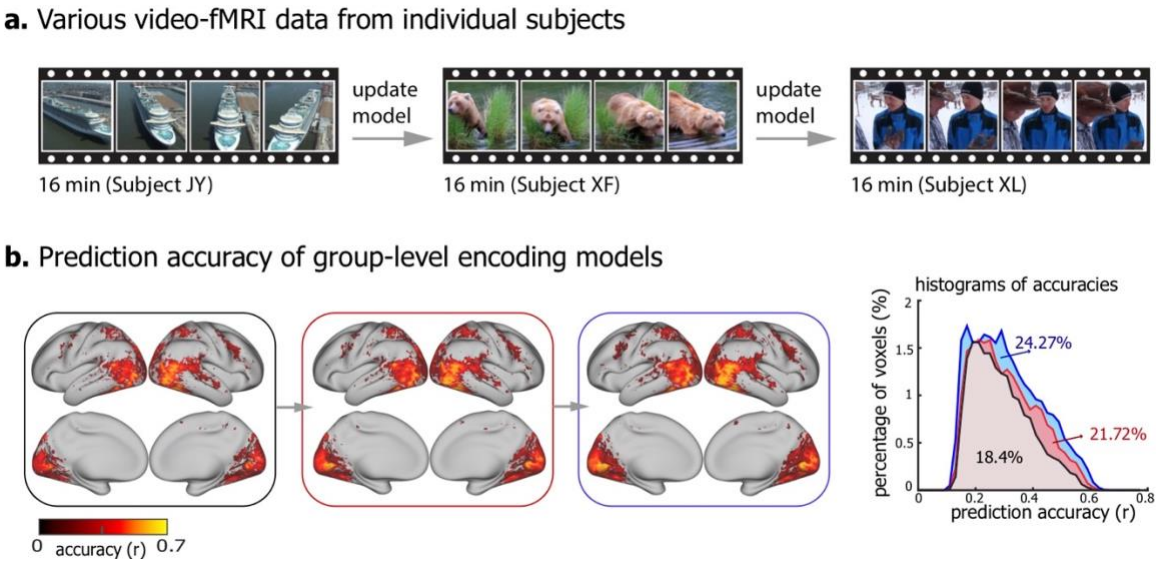

Figure 4.6 **Group-level encoding models.** (a) Distinct video-fMRI dataset obtained from different subjects when watching different natural videos. (b) The voxel-wise prediction accuracy of group-level encoding models before and after every incremental update (permutation test, corrected at FDR $q<0.01$). The right is the histograms of prediction accuracies of incrementally updated encoding models. The colored numbers are the total percentages of predictable voxels. The testing accuracy is averaged across three subjects.

Table 4.1 Online learning algorithm for training population-based encoding models.

Algorithm 1: Online learning algorithm for training population-based encoding models

1: $\mathbf{G}^0 \leftarrow \mathbf{0},\ \boldsymbol{w}_v^0 \leftarrow \mathbf{0},\ n^0 \leftarrow 0,\ \lambda^0 = 0$

2: **While** new data* is available: $\mathbf{X},\ \boldsymbol{r}_v^1,\ n^1$

3: $\theta = \frac{n^1}{n^0+n^1}$

4: $\mathbf{F}^1 = \mathbf{DimensionReduction}\big(\mathbf{ResNet}(\mathbf{X})\big)$

5: $\mathbf{G}^1 = [\mathbf{F}^1]^{\mathrm{T}}\mathbf{F}^1/n^1$

6: $\mathbf{G} = (1-\theta)\mathbf{G}^0 + \theta\mathbf{G}^1$

7: $\boldsymbol{w}_v = (1-\theta)(\mathbf{G}+\lambda\mathbf{I})^{-1}(\mathbf{G}^0+\lambda^0\mathbf{I})\boldsymbol{w}_v^0 + \theta(\mathbf{G}+\lambda\mathbf{I})^{-1}\,[\mathbf{F}^1]^{\mathrm{T}}\boldsymbol{r}_v^1/n^1$ with cross validation

8: $\mathbf{G}^0 \leftarrow \mathbf{G},\ \boldsymbol{w}_v^0 \leftarrow \boldsymbol{w}_v, n^0 \leftarrow n^0+n^1, \lambda^0 = \lambda$

9: **Output**: $\boldsymbol{w}_v$

* $\mathbf{X}$ is the new visual stimuli, $\boldsymbol{r}_v^1$ is the cortical response, and n^1 is the number of samples

5. DEEP PREDICTIVE CODING NETWORK FOR OBJECT RECOGNITION

*Modified and formatted for dissertation from the article under review in *ICML* [39]

5.1 Introduction

There are mExperiment There are mExperiment Convolutional neural networks (CNN) have achieved great success in image recognition. Classical CNN models, e.g. AlexNet [19], VGG [20], GoogLeNet [6], ResNet [21], SENets [177], NASNet [178], have improved the performance in computer vision, while these models generally become deeper and wider by using more layers [6, 20, 21] or/and filters [6, 179]. Despite various ways of architectural reconfiguration, these models all scale up from the same principle of computation: extracting image features by a feedforward pass through stacks of convolutional layers.

Although it is inspired by hierarchical processing in biological visual systems [180], CNN differs from the brain in many aspects. Unlike CNN, the brain achieves robust visual perception by using feedforward, feedback and recurrent connections [181, 182]. Information is processed not only through a bottom-up pathway running from lower to higher visual areas, but also through a top-down pathway running in the opposite direction. Such bi-directional processes enable humans to perform a wide range of visual tasks, including object recognition. For human vision, feedforward processing is essential to rapid recognition [46, 89], e.g. when visual input is too brief to recruit feedback and recurrent processing [97]. However, feedback processing improves object recognition and enables cognitive processes to influence perception [183, 184]. In neuroscience, the interplay between feedforward and feedback processes is described by hierarchical *predictive coding* [13, 14, 16, 36-38]. It states that the feedback connections from a higher visual area to a lower visual area carry predictions of lower-level neural activities; feedforward connections carry the errors between the predictions and the actual lower-level activities. As a result, the brain dynamically updates its representations to progressively refine its perceptual and behavioral decisions.

Inspired by this brain theory, we designed a bi-directional and recurrent neural net (i.e. PCN). Given image input to PCN, it runs recursive cycles of bottom-up and top-down computation

to update its internal representations towards minimization of the residual error between bottom-up input and top-down prediction at every layer in the network. Using predictive coding as its computational mechanism, PCN differs from feedforward-only CNNs that currently dominate computer vision. It is a model with dynamics that uses recursive and bi-directional computation to extract better representations of the input such that the input is predictable by the extracted representation. When it is unfolded in time, PCN runs a longer cascade of nonlinear transformations by running more cycles of bottom-up and top-down computation through the same architecture without adding more layers, units, or connections.

To explore its value, we designed PCN with convolutional layers stacked in both feedforward and feedback directions. We trained and tested PCN for image classification with benchmark datasets: CIFAR-10 [185], CIFAR-100 [185], SVHN [186], and MNIST [187]. Our focus was to explore the intrinsic advantages of PCN over its feedforward-only counterpart: a plain CNN model without feedback connection or any mechanism for recurrent dynamics. It turned out that PCN always outperformed the plain CNN model, and its accuracy tended to improve given more cycles of computation over time. Relative to the classical models, PCN yielded competitive performance in all benchmark tests despite much less layers in PCN. As we did not attempt to optimize the performance by trying many learning parameters or model architectures, there is much room for future studies to further improve or extend the model on the basis of a similar notion.

5.2 Related Work

Current progress in computer vision is more driven by engineering goals as opposed to inspiration from the brain [188]. Findings from recent studies demonstrate that deep convolutional neural networks use representations similar to those in the brain [7-10, 22, 30]. However, many gaps are yet to be filled to bridge biological and artificial visual systems. A biologically plausible model of vision should take into account feedback and recurrent connections, which are abundant in primate brains [181, 182]. A limited number of studies have taken on this direction from the perspective of computational neuroscience or computer vision.

O'Reilly et al. demonstrated that feedback connections could enable top-down representations to fill incomplete bottom-up representations to improve recognition of partially occluded objects [189]. Exploiting a similar idea, Spoerer et al. built a recurrent CNN (with 2

hidden layers) using feedforward, feedback, and lateral connections to enable recurrent processing that dynamically updated the internal representations as the sum of bottom-up, top-down, and lateral contributions [190]. Trained and tested with synthesized images of digits, their recurrent CNN yielded more robust recognition of digits in cluttered and occluded images. However, that model did not embody an explicit computational mechanism to ensure recurrent processing dynamics to converge over time. Although compelling from the neuroscience perspective, the models in the above studies were relatively simple and shallow, and they were not tested in naturalistic visual scenarios of primary interest to computer vision.

In computer vision, Liang et al. added recurrent connections into each layer of a feedforward CNN to allow the activity of each unit to be modulated by activities of its neighboring units within the same layer [191]. Although it was inspired by contextual modulation in biological vision, this model did not account for feedback connections, which are abundant in the brain. Stollenga et al. added feedback connections to a trained CNN to enable attentional selection of filters for the model to achieve better object classification [192]. Recently, Canziani et al. built a bi-directional model with a feedforward discriminant subnet, a feedback generative subnet, as well as lateral connections to bridge the two subnets; training the model for video prediction helped the model yield more stable object recognition given video input [193]. These studies described above highlight the roles of feedback and/or recurrent processes in computing or learning better representations than models with only feedforward processes. What remains unresolved is a biologically plausible mechanism that allows feedforward, feedback, and recurrent processes to interact with one another in order for the model to manifest internal dynamics that support various learning objectives.

In this regard, we may seek inspiration from the brain. Predictive coding is an influential theory of neural processing in vision and beyond [15, 37, 38] as supported by empirical evidence [194-198]. In a seminal paper [199], Rao and Ballard postulated that the brain learns a hierarchical internal model of the visual world. Each level in this model attempts to predict the responses at its lower level via feedback connections; the error between this prediction and the actual response is sent to the higher level via feedforward connections. Friston et al. further generalized this notion into a unified brain theory for perception and action [200]. Chalasani et al. used predictive coding to train a deep neural net to learn a hierarchy of sparse representations of data without supervision [201]. Lotter et al. explored video prediction as an unsupervised learning objective based on

predictive coding [202]; however the model trained in this way may not be able to learn sufficiently abstract representation to support such tasks as object recognition. Spratling et al. explored the use of predictive coding for object recognition; however, their model was limited a shallow network architecture for much simplified scenarios [203].

Inspired by but different from models in prior studies [16, 203, 204], a hierarchical, bidirectional, and recurrent model is proposed and implemented herein as a brain-inspired model for computer vision. This model operates with the theory of predictive coding to generate dynamic internal representations by recursive bottom-up and top-down computation via feedforward and feedback connections across cascaded layers in a deep hierarchy, and recurrent connections to convey information over time within the same layer. The internal representations are updated to progressively reduce the error of top-down prediction of lower-level representations, while the prediction errors are conveyed upward to higher levels. To train this network, the representations at the highest level, after multiple cycles of recursive updating, are used to classify the input image. With labeled images, the model parameters are trained through backpropagation in time and across layers. In this study, we trained and tested such a deep predictive coding network (PCN) with several datasets: CIFAR-10, CIFAR-100, SVHN, and MNIST.

5.3 Methods

5.3.1 Predictive coding

Central to the theory of predictive coding is that the brain continuously generates top-down predictions of bottom-up inputs. The representation at a higher level predicts the representation at its lower level. The difference between the predicted and actual representation elicits an error of prediction, and propagates to the higher level to update its representation towards improved prediction. This repeats throughout the hierarchy until the errors of prediction diminish, or the bottom-up process no longer conveys any "new" (or unpredicted) information to update the hidden representation. Thus, predictive coding is a computational mechanism for the model to recursively update its internal representations of an image towards convergence.

In the following mathematical description of this dynamic process in PCN, italic lowercase letters are used as symbols for *scalars*, bold lowercase letters for column **vectors**, and bold uppercase letters for **MATRICES**. The representation at layer l and time t is denoted as $\mathbf{r}_l(t)$. The

weights of feedforward connections from layer l-1 to layer l are denoted as $\mathbf{W}_{l-1,l}$. The weights of feedback connections from layer l to layer l-1 are denoted as $\mathbf{W}_{l,l-1}$.

In PCN, the higher-level representation, $\mathbf{r}_l(t)$, predicts its lower-level representation as $\mathbf{p}_{l-1}(t)$ via linear weighting $\mathbf{W}_{l,l-1}$, as shown in Eq. (1). The prediction error, $\mathbf{e}_{l-1}(t)$, is the difference between $\mathbf{p}_{l-1}(t)$ and $\mathbf{r}_{l-1}(t)$ as in Eq. (2).

$$\mathbf{p}_{l-1}(t) = \left(\mathbf{W}_{l,l-1}\right)^{\mathrm{T}}\mathbf{r}_l(t) \qquad (1)$$

$$\mathbf{e}_{l-1}(t) = \mathbf{r}_{l-1}(t) - \mathbf{p}_{l-1}(\mathrm{t}) \qquad (2)$$

5.3.1.1 Feedforward process

For the feedforward process, the prediction error at layer l-1, $\mathbf{e}_{l\text{-}1}(t)$, propagates to the upper layer l to update its representation, $\mathbf{r}_l(t)$, so the updated representation reduces the prediction error. To minimize $\mathbf{e}_{l\text{-}1}(t)$, let's define a loss as the sum of the squared errors normalized by the variance of the representation, $\sigma_{l\text{-}1}^2$, as in Eq. (3).

$$e_{l\text{-}1}(t) = \frac{1}{\sigma_{l\text{-}1}^2}\|\mathbf{e}_{l\text{-}1}(t)\|_2^2 \qquad (3)$$

The gradient of $e_{l\text{-}1}(t)$ with respect to $\mathbf{r}_l(t)$ is as Eq. (4).

$$\frac{\partial e_{l-1}(t)}{\partial \mathbf{r}_l(t)} = -\frac{2}{\sigma_{l-1}^2}\mathbf{W}_{l,l-1}\mathbf{e}_{l-1}(t) \qquad (4)$$

To minimize $e_{l\text{-}1}(t)$, $\mathbf{r}_l(t)$ is updated by gradient descent with an updating rate, α_l, as shown in Eq. (5).

$$\mathbf{r}_l(t+1) = \mathbf{r}_l(t) - \alpha_l\left(\frac{\partial e_{l-1}(t)}{\partial \mathbf{r}_l(t)}\right) = \mathbf{r}_l(t) + \frac{2\alpha_l}{\sigma_{l-1}^2}\mathbf{W}_{l,l-1}\mathbf{e}_{l-1}(t) \qquad (5)$$

If the weights of feedback connections are the transpose of those of feedforward connections $\mathbf{W}_{l,l-1} = \left(\mathbf{W}_{l-1,l}\right)^{\mathrm{T}}$, the update rule in Eq. (5) can be rewritten as a *feedforward* operation, as in Eq. (6).

$$\mathbf{r}_l(t+1) = \mathbf{r}_l(t) + a_l\left(\mathbf{W}_{l-1,l}\right)^{\mathrm{T}}\mathbf{e}_{l-1}(t) \qquad (6)$$

where the last term indicates forwarding the prediction error from layer l-1 to layer l to update the representation with an updating rate $a_l = \frac{2\alpha_l}{\sigma_{l-1}^2}$.

5.3.1.2 Feedback process

For the feedback process, the top-down prediction is used to update the representation at layer l, $\mathbf{r}_l(t)$, to reduce the prediction error $\mathbf{e}_l(t)$. Similar to feedforward process, the error is

minimized by gradient descent, where the gradient of $e_l(t)$ with respect to $\mathbf{r}_l(t)$ is as Eq. (7), and $\mathbf{r}_l(t)$ is updated with an updating rate β_l as shown in Eq. (8).

$$\frac{\partial e_l(t)}{\partial \mathbf{r}_l(t)} = \frac{2}{\sigma_l^2}\left(\mathbf{r}_l(t) - \mathbf{p}_l(\mathrm{t})\right) \qquad (7)$$

$$\mathbf{r}_l(t+1) = \mathbf{r}_l(t) - \beta_l\left(\frac{\partial e_l(t)}{\partial \mathbf{r}_l(t)}\right) = \left(1 - \frac{2\beta_l}{\sigma_l^2}\right)\mathbf{r}_l(t) + \frac{2\beta_l}{\sigma_l^2}\mathbf{p}_l(t) \qquad (8)$$

Let $b_l = \frac{2\beta_l}{\sigma_l^2}$ and Eq. (8) is rewritten as follows.

$$\mathbf{r}_l(t+1) = (1 - b_l)\mathbf{r}_l(t) + b_l\mathbf{p}_l(t) \qquad (9)$$

E. (9) reflects a *feedback* process that the representation at the higher layer, $\mathbf{r}_{l+1}(t)$, generates a top-down prediction, $\mathbf{p}_l(t)$, and influences the representation at the lower level, $\mathbf{r}_l(t)$.

5.3.1.3 Nonlinearity

To add nonlinearity to the above feedforward and feedback processes, a nonlinear activation function is applied to the output of each convolutional layer (except the input layer, i.e. $l = 0$). A rectified linear unit (ReLU) [205] converts Eqs. (6) and (9) to nonlinear processes as below.

Nonlinear feedforward process:

$$\mathbf{r}_l(t+1) = \mathrm{ReLU}\left(\mathbf{r}_l(t) + a_l\left(\mathbf{W}_{l-1,l}\right)^{\mathrm{T}}\mathbf{e}_{l-1}(t)\right) \qquad (10)$$

Nonlinear feedback process:

$$\mathbf{r}_l(t+1) = \mathrm{ReLU}\left((1 - b_l)\mathbf{r}_l(t) + b_l\mathbf{p}_l(t)\right) \qquad (11)$$

5.3.2 Network architecture

We implemented this algorithm in several PCNs, all of which included convolutional layers stacked in both feedforward and feedback directions and recurrent connections within each layer as shown in Fig. 5.1a. These PCNs were trained and tested for object recognition with four benchmark datasets: CIFAR-10, CIFAR-100, SVHN and MNIST. For comparison, several feedforward-only CNNs were built with the same architecture as the feedforward pathway in corresponding PCNs, and were trained and tested with the same datasets. We refer to these feedforward-only CNNs as the plain networks, from which the PCNs were built upon by adding feedback and recurrent connections for dynamic processing.

Plain CNN models. The architecture of our plain CNN models were similar to the architecture of VGG nets [20]. Briefly, the basic architecture included 6 or 8 convolutional layers

and 1 classification layer. All convolutional layers used 3×3 filters but different numbers of filters, and used rectified linear unit (ReLU) as the nonlinear activation function. For some layers where the number of filters is doubled, the feature maps were reduced by applying 2×2 max-pooling with a stride of 2 after convolution. Batch normalization [206] was not used. The classification layer included global average pooling and a fully-connected (FC) layer followed by softmax. On the basis of this architecture, we built 5 VGG-like models that varied in the number of layers and filters, and trained and tested the models with 4 datasets. Table 5.1 summarizes the architecture of each model.

Predictive coding network (PCN). Starting from each of the plain CNN models, we added feedback and recurrent connections to form a corresponding PCN. Fig. 5.1a shows a 9-layer PCN, running recursive bottom-up and top-down processing based on predictive coding. In PCN, feedback connections from one layer to its lower layer were constrained to be the transposed convolution [207] which is the transpose of the feedforward counterparts, setting apart our models from those in related work on predictive coding [16, 203, 204]. As such, both feedforward and feedback connections encoded spatial filters. The former was applied to the errors of the top-down prediction of lower-level representation; the latter was applied to high-level representation in order to predict the lower-level representation. As in the brain, feedforward and feedback connections were reciprocal in PCN. The weights of feedback connections had the identical dimension as the transposed weights of feedforward connections. For layers where max-pooling was applied after feedforward convolution, bilinear unsampling was applied before feedback convolution to ensure that the dimension of top-down prediction could match the dimension of lower-level representation.

An optional constraint to PCN was to use the same set of weights for both feedforward and feedback connections as in some prior studies [16, 203, 204]. In other words, the weights of feedback connections were the transposed weights of feedforward connections. With this weight sharing, top-down predictions via feedback connections tended to approach lower-level representations. The PCN would have the same number of parameters as the corresponding plain model. Without this optional constraint of weight sharing, feedforward and feedback weights were assumed to be independent.

5.3.3 Recursive computation

Unlike feedforward-only networks, PCN runs a dynamic process to update its internal representation throughout the hierarchy (Fig. 5.1.b). Given an input image, PCN first runs through the feedforward path from the input layer to the last convolutional layer at $\boldsymbol{t} = \mathbf{0}$, equivalent to a plain CNN model. For $\boldsymbol{t} = \mathbf{1}$, PCN first runs a feedback process and then a feedforward process to update the representations in the hierarchy. In the feedback process, the representation at each layer is updated by a top-down prediction from the higher layer according to Eq. (11). The feedback process runs from the highest convolutional layer to the input layer. In the feedforward process, the representation at each layer is updated by a bottom-up error according to Eq. (10). This procedure is repeated over time. After some cycles, the representation is used as the input to the classification layer to classify the image (see **Algorithm 2** in Table 5.5).

5.3.4 Model training

When PCN is trained for image classification, the error backpropagates across layers and in time to update the model parameters. The update rates are constrained to be non-negative by using ReLU, and are learnable parameters specific to each filter in each layer.

We evaluated two types of PCNs with regard to an optional constraint: the feedforward and feedback connections share the same convolutional weights. With this weight sharing, the feedforward operation and the feedback operation use the same weights. Without the constraint, the feedforward and feedback weights are initialized interpedently.

In this work, we evaluated these two types of PCNs with a varying number of recursive cycles ($t = 0, 1, 2, \cdots, 6$) and with different model architectures (labeled as A through E in Table 5.1). We use *Plain-A* to represent the plain network with architecture A, and use *PCN-A-t* to represent the PCN with architecture A and t cycles of recursive computation. *PCN-A-t (tied)* and *PCN-A-t* represent the PCNs *with* and *without* weight sharing, respectively.

We used PyTorch [208] to implement, train, and test the models described above. The convolutional weights and linear weights were initialized to be uniformly random (the default setting in PyTorch). The feedforward and feedback update rates were initialized as 1.0 and 0.5, respectively. The models were trained using mini-batches of a size 128 and without using dropout regularization [71].

5.4 Experiments

Methods and MaterialsWe trained and tested PCN for image classification with data in CIFAR-10/100, SVHN and MNIST, in comparison with plain CNN using the same feedforward architecture. With random initialization, PCN (or CNN) was trained for 5 times; the best and mean±std top-1 accuracy was reported as below.

5.4.1 CIFAR-10 and CIFAR-100

The CIFAR-10/100 dataset includes 50,000 training images and 10,000 testing images in 10 or 100 object categories. Each image is a 32×32 RGB image. PCN (or CNN) were trained on the training set and evaluated on the test set. All images were normalized per channel (i.e. subtract the mean and divide by the standard deviation). For training, we used translation and horizontal flipping for data augmentation. We used stochastic gradient decent to train PCN (or CNN) with a weight decay of 0.0005 and a momentum of 0.9. The learning rate was initialized as 0.01 and was divided by 10 when the error reached the plateau after training for 80, 140, 200 epochs. We stopped after 250 epochs. The hyper-parameters for learning were set based on validation with 10,000 images in the training set.

5.4.1.1 PCN vs. CNN

During training, PCN converged much faster than its CNN counterpart (Fig. 5.2, top), especially when feedforward and feedback connections did not share weights. With testing data, PCN also yielded better accuracy than the plain CNN model (Fig. 5.2, bottom). For example, *PCN improved* the classification accuracy from 62.11% to 72.48% on CIFAR-100, relative to the plain CNN model. See Table 5.2 for more results for comparison with other classical or state-of-the-art models. Without being pushed for high accuracy, PCN showed a similar accuracy as ResNet [21], but relatively lower than the pre-activation ResNet (Pre-act-ResNet) [209] or the wide residual network (WRN) [179], which used a much deeper or much wider architecture than the models explored in this study.

5.4.1.2 PCN with different recursive cycles

The accuracy of PCN depended on the number of cycles that recursively updated its internal representations. Fig. 5.3 shows that the accuracy of PCN tended to increase given more cycles of computation, especially if feedforward and feedback processes did not share the same weights.

To understand why this was the case, we looked into some testing images that were mis-classified by CNN but not by PCN. At each time step (0 through 6), PCN computed a different representation of an image that yielded a different probability distribution across different categories (Fig. 5.4). Classification was less definitive and/or inaccurate at early time steps. At later time steps, the network corrected itself to yield more definitive and accurate classification. It was true especially for ambiguous images, where a cat looked like a dog, or a deer looked like a horse, even for humans. See more examples in Fig. 5.4.

5.4.1.3 Generative prediction in PCN

When it was trained for image classification, PCN was not explicitly optimized to reconstruct the input image, unlike a previous work that used video prediction as the learning objective [202]. Nevertheless, the top-down process in PCN was able to reconstruct the input with high accuracy. Although this was expected for PCN with weight sharing, reconstruction was also reasonable even for PCN without weight sharing (Fig. 5.5). This result was surprising, and implied that PCN, without any architectural constraint to enable image reconstruction, is able to reshape itself to predict or reconstruct the input, even when it is trained for a discriminative task, e.g. object recognition. Speculatively PCN potentially provides a new way to simultaneously train a discriminative network for object recognition and a generative network for prediction or reconstruction.

5.4.2 SVHN

SVHN is a dataset of Google's Street View House Numbers images [186] and contains more than 600,000 color images of size 32×32, divided into training set, testing set and an extra set. The task of this dataset is to classify the digit located at the center of each image. Since the task is easier than CIFAR datasets, we implemented PCN with simpler network architectures (see Table 5.1). To validate the hyper parameters, we randomly selected 400 samples per class from the training set and 200 samples per class from the extra set for validation, as in [210]. The remainder of the training set and the extra set were used for training. The preprocessing for SVHN was the same as for CIFAR, i.e. per-channel normalization. No data augmentation was used. We used the Adam [211] optimization with a weight decay of 0.0005 and an initial learning rate of 0.001 for a 20-10-10 epoch schedule. The exponential decay rates for the first and second moment estimates were 0.9 and 0.99, respectively. Table 5.3 shows the classification performance for this

dataset. Like what we found for the CIFAR dataset, PCN always outperformed the plain CNN counterpart.

5.4.3 MNIST

The MNIST dataset consists of hand written digits 0-9. There are 60,000 training images and 10,000 testing images in total. Each image is a gray image of size 28x28. For this dataset, the same network architecture as used for SVHN is adopted. The training procedure was the same as for SVHN. Table 5.4 shows the classification performance for this dataset. PCN consistently performed better than its CNN counterpart. The best PCN achieves 0.36% error rate, comparable to some previous state-of-the-art models.

5.5 Discussion and Conclusion

What defines PCN are 1) the use of bi-directional and recurrent connections as opposed to feedforward-only connections, and 2) the use of predictive coding as a mechanism for the model to recursively run bottom-up and top-down processes. When it is trained for image classification, the model dynamically refines its representation of the input image towards more accurate and definitive recognition. As this computation is unfolded in time, PCN reuses a single architecture and the same set of parameters to run an increasingly longer cascade of nonlinear transformation.

We say it is "longer" instead of "deeper", because the notion behind PCN is different from the mindset in deep learning that more layers are required to model more complex and nonlinear relationships in data. Making a model increasingly deeper is arguably less efficient or scalable, bringing a set of challenges or burdens, e.g. the need for more computational resource and training data. In contrast, the brain does not use a deeper network to do more challenging tasks. A more challenging task simply takes the brain longer time to process information through the same network.

Predictive coding tells PCN how to compute but not how to learn. In this study, PCN is trained for image classification based on the representation emerging from the top layer after multiple cycles of computation. The error of classification backpropagates (top-down and bottom-up) across layers and in time to update the model parameters for multiple times (as many as the cycles of recursive computation) per training example or batch of examples. This helps the learning to converge faster, while utilizing full knowledge in training data. If an image takes the model

more cycles of computation to converge its representation, it means that the image has more information than what the model can explain or generate, and thus the image carries a greater value for the model to learn. Therefore, it is more desirable to train PCN for more challenging visual tasks, e.g. images that are ambiguous or difficult to recognize, while reducing the need for a large number of otherwise "simple" training examples.

For image classification, PCN takes an image as the input for all cycles of its recursive computation, while the errors of top-down prediction sent to the first hidden layer vary across cycles or in time. When the input is not a static image but a video, the input to the first hidden layer represents the errors of prediction of the present video frame given the model's representations from the past frames. This would enable the model to compute and learn representations of both spatial and temporal information in videos, which is an important aspect that awaits to be explored in future studies.

As an initial step to explore predictive coding in computer vision, it was our intention to start and compare with models with a basic CNN architecture (like that of VGG) in order to focus on evaluation of the value of using predictive coding as a computational mechanism. However, we expect that some network modules are readily applicable to PCN as well as CNN, including batch normalization [206] and short-cut connections [21]. In addition, the update rates for top-down and bottom-up computation may be trainable as time-variant parameters as opposed to constants assumed in the current implementation. Augmentation of training data or regularization techniques, e.g. dropout [71] may also help to improve the model's performance in image classification. In future studies, we will explore alternative architectures and learning strategies for larger and more training images, e.g. ImageNet [19].

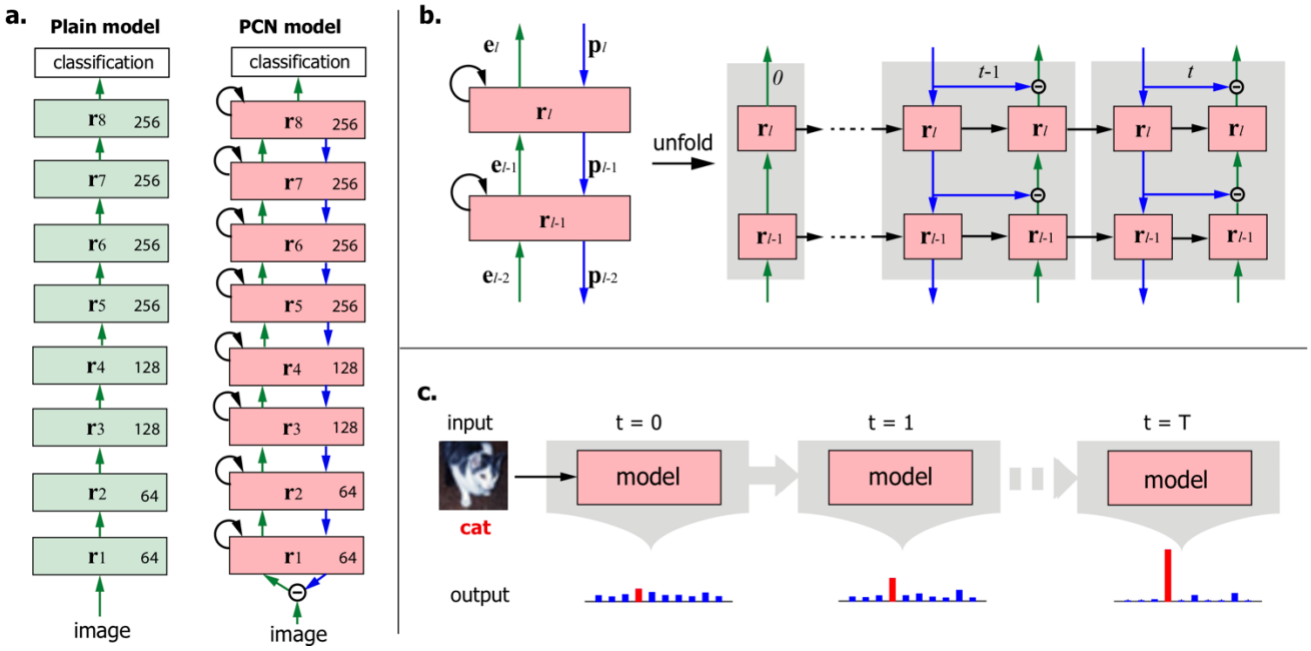

Figure 5.1 a) An example PCN with 9 layers and its feedforward-only CNN (or the plain model). b) Two-layer substructure of PCN. Feedback (blue), feedforward [149], and recurrent (black) connections convey the top-down prediction, the bottom-up prediction error, and the past information, respectively. c) The dynamic process in the PCN iteratively updates and refines the representation of visual input over time. PCN outputs the probability over candidate categories for object recognition. The bar height indicates the probability and the red indicates the ground truth.

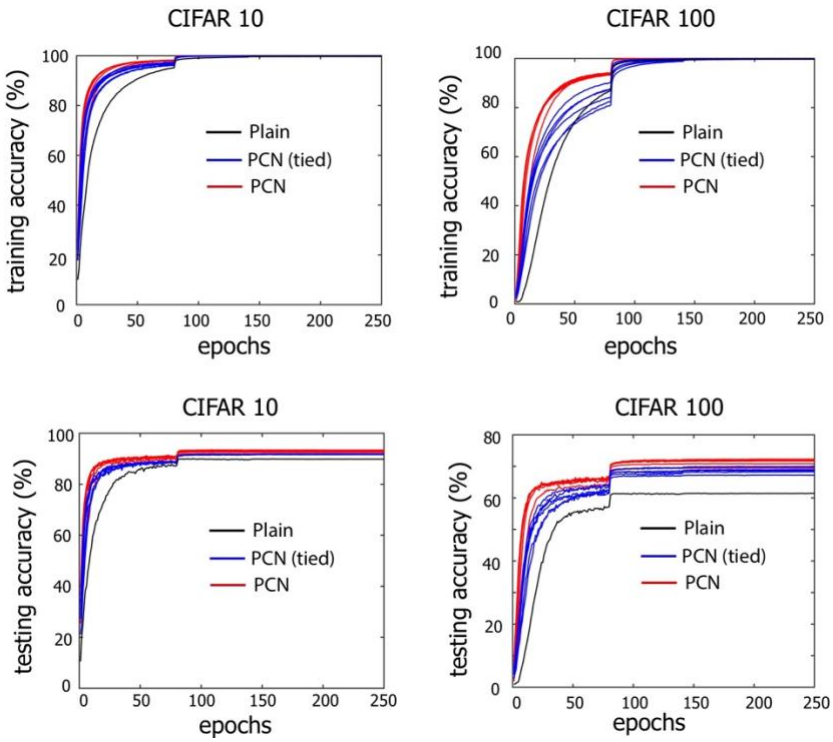

Figure 5.2 Training (top) and testing (bottom) accuracies for PCN vs. CNN with matched feedforward architectures for training with CIFAR-10 (left) and CIFAR-100 (right). Each curve represents the average over 5 repeats of one model with different cycles of recursive computation, ranging from 1 to 6.

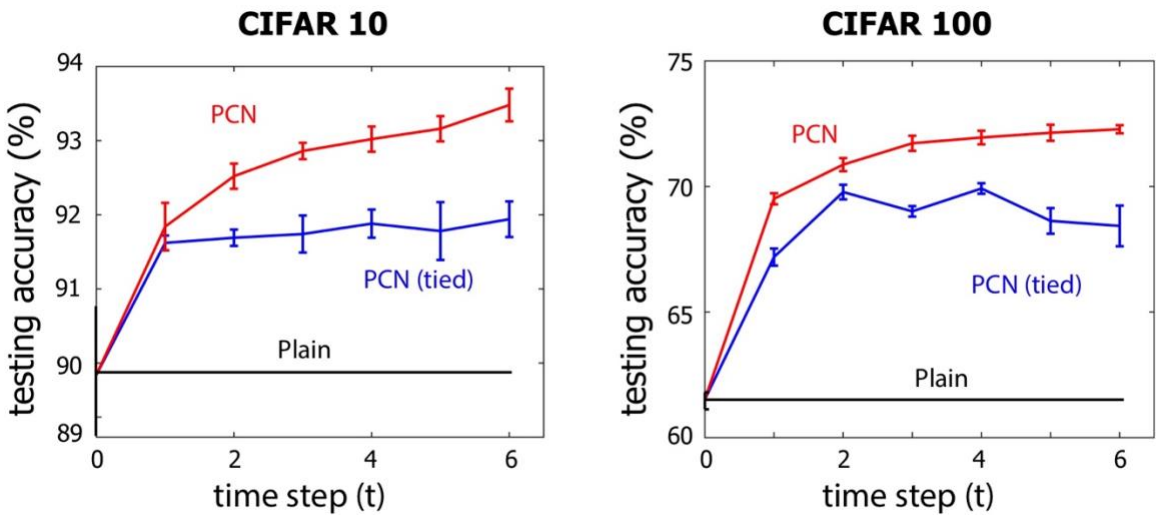

Figure 5.3 Testing accuracies of PCNs with different time steps.

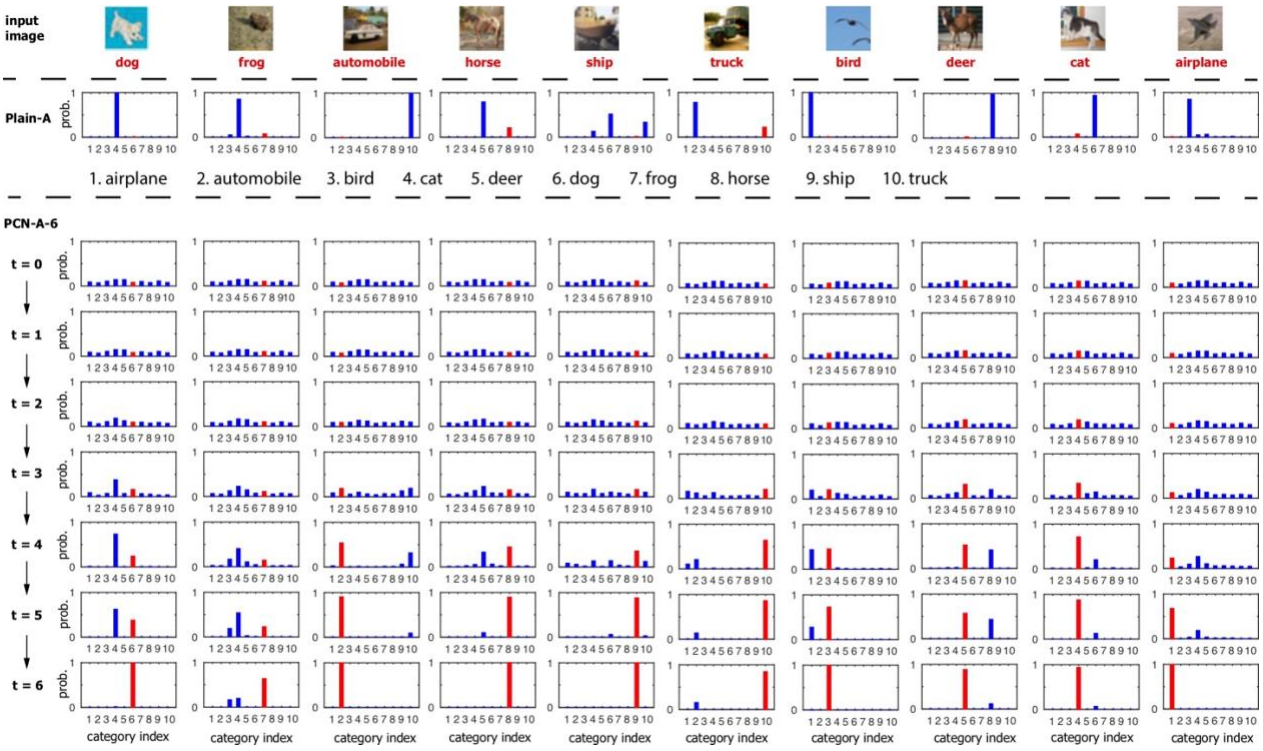

Figure 5.4 Image classification at different time steps for PCN-A-6 (bottom) in comparison with the plain CNN model (middle) for each of the 10 testing images misclassified by CNN (Plain-A). Each plot shows the probabilities over 10 classes in CIFAR-10. The red represents the ground truth.

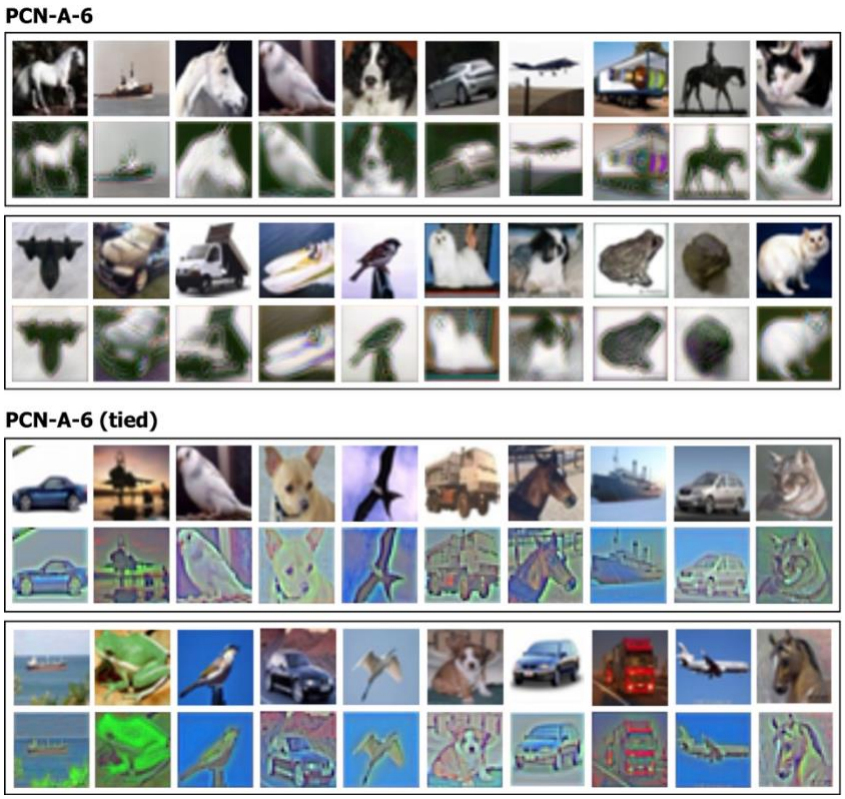

Figure 5.5 Top-down image prediction by PCN. Here shows example testing images in CIFAR-10 and their corresponding images predicted by PCNs.

Table 5.1 Architectures for PCN. Each column is a model. The layers with the same color have the same feature map size.

CIFAR-10/100			**SVHN/ MNIST**	
A	B	C	D	E
9 layers	9 layers	7 layers	7 layers	7 layers
input image				
conv3 -**64**	conv3 -**32**	conv3 -**32**	conv3 -**32**	conv3 -**16**
conv3 -**64**	conv3 -**32**	conv3 -**32**	conv3 -**32**	conv3 -**16**
conv3 -**128**	conv3 -**64**	conv3 -**64**	conv3 -**64**	conv3 -**32**
conv3 -**128**	conv3 -**64**	conv3 -**64**	conv3 -**64**	conv3 -**32**
conv3 -**256**	conv3 -**128**	conv3 -**128**	conv3-**128**	conv3 -**64**
conv3 -**256**	conv3 -**128**	conv3 -**128**	conv3-**128**	conv3 -**64**
conv3 -**256**	conv3 -**128**			
conv3 -**256**	conv3 -**128**			
global average pooling, FC-10/100, softmax				

Table 5.2 Compare PCNs with start-of-the-art models on CIFAR-10/100 datasets. #Layer and #Parameter are the number of layers and parameters, respectively.

Models			**CIFAR10/100**	
Methods	#Layer	#Parameter	Accuracy (%)	
Maxout[210]	-	-	90.62	61.43
dasNet [192]	-	-	90.78	66.22
NIN [212]	-	-	91.19	64.32
DSN [213]	-	-	91.78	65.43
RCNN [191]	6	1.86M	92.91	68.25
FitNet [214]	19	2.5M	91.61	64.96
Highway[215]	19	2.3M	92.46	67.76
ResNet [21]	110	1.7M	93.57	-
	164	1.7M	-	74.84
	1001	10.2M	-	72.18
	1202	19.4M	92.07	-
Pre-act-ResNet [209]	110	1.7M	93.63	-
	164	1.7M	94.54	75.67
	1001	10.2M	95.08	77.29
WRN-40-4	40	8.9M	95.47	78.82
WRN-16-8	16	11M	95.73	79.57
WRN-28-10 [179]	28	36.5M	96.00	80.75
DenseNet [216]	250	15.3M	96.28	82.40
Plain-A	9	2.33M	90.61	62.11
PCN-A-6 (tied)	9	2.33M	92.26	69.44
PCN-A-6	9	4.65M	**93.83**	**72.58**
Plain-B	9	0.58M	89.53	62.21
PCN-B-2 (tied)	9	0.58M	90.76	65.57
PCN-B-6	9	1.16M	92.80	69.34
Plain-C	7	0.29M	88.23	61.36
PCN-C-2 (tied)	7	0.29M	89.56	64.09
PCN-C-6	7	0.57M	92.40	68.31

Table 5.3 Compare PCNs with start-of-the-art models on SVHN. The accuracy was obtained from five repeats.

SVHN			
Methods	#Layer	#Parameter	error rate (%)
Maxout[210]	-	-	2.47
NIN [212]	-	-	2.35
Stochastic pooling [217]	-	-	2.80
Dropconnect [218]	-	-	1.94
DSN [213]	-	-	1.92
RCNN [191]	6	2.67M	1.77
FitNet [214]	13	1.5M	2.42
WRN-16-8 [179]	16	11M	1.54
Plain-D	7	0.29M	3.21(3.41±0.13)
PCN-D-2 (tied)	7	0.29M	2.63(2.92±0.11)
PCN-D-6	7	0.57M	2.28(2.42±0.09)
Plain-E	7	0.07M	3.19(3.41±0.13)
PCN-E-1 (tied)	7	0.07M	2.74(2.91±0.11)
PCN-E-6	7	0.14M	**2.24**(2.42±0.10)

Table 5.4 Compare PCNs with the start-of-the-art models on MNIST. The accuracy was obtained from five repeats.

MNIST			
Methods	#Layer	#Parameter	error rate (%)
Maxout[210]	-	-	0.45
NIN [212]	-	-	0.47
Stochastic pooling [217]	-	-	0.47
Dropconnect [218]	-	-	0.21
DSN [213]	-	-	0.39
RCNN [191]	6	0.67M	0.31
FitNet [214]	-	-	0.51
Hierarchical PC/BC-DIM [203]	-	-	2.19
Plain-D	7	0.29M	0.53(0.59±0.04)
PCN-D-1 (tied)	7	0.29M	0.43(0.50±0.06)
PCN-D-1	7	0.57M	0.38(0.46±0.06)
Plain-E	7	0.07M	0.68(0.74±0.03)
PCN-E-1 (tied)	7	0.07M	0.43(0.51±0.06)
PCN-E-4	7	0.14M	**0.36**(0.48±0.06)

Table 5.5 Algorithm of the Deep Predictive Coding Network.

Algorithm 2 Deep Predictive Coding Network

Input static image: **x**

2. $\mathbf{r}_0(t) \leftarrow \mathbf{x}$
3.
4. **for** l = 0 **to** L-1 **do**
5. $\mathbf{r}_{l+1}(0) \leftarrow \mathrm{ReLU}\left(\mathrm{FFConv}\left(\mathbf{r}_l(0)\right)\right)$
6.
7. **for** t = 1 to T **do**
8. **for** l = L **to** 1 **do**
9. $\mathbf{p}_{l-1}(t\text{-}1) \leftarrow \mathrm{FBConv}\left(\mathbf{r}_l(t\text{-}1)\right)$
10. **if** l > 1 **do**
11. $\mathbf{r}_{l-1}(t\text{-}1) \leftarrow \mathrm{ReLU}\left((1\text{-}b)\mathbf{r}_{l\text{-}1}(t\text{-}1)+b\mathbf{p}_{l\text{-}1}(t\text{-}1)\right)$
12. **for** l = 0 **to** L-1 **do**
13. $\mathbf{e}_l(t) \leftarrow \mathbf{r}_l(t) - \mathbf{p}_l(t\text{-}1)$
14. $\mathbf{r}_{l+1}(t) \leftarrow \mathrm{ReLU}\left(\mathbf{r}_{l+1}(t\text{-}1)+a\mathrm{FFConv}\left(\mathbf{e}_l(t)\right)\right)$
15.
16. output $\mathbf{r}_{\mathrm{L}}(\mathrm{T})$ for classification

Note: *FFConv* represents the feedforward convolution, *FBConv* represents the feedback convolution.

6. SUMMARY

Functional imaging for vision has been mostly limited to mapping either low-level visual elements (e.g. orientation or color) [219] or high-level object categories [60, 100]. However, little is known about how middle-level features are represented [8, 9, 44, 53] or how different levels of features are related to one another through neural computation. In this dissertation, I established the experimental and analysis techniques for using fMRI to map cortical representations of all levels of visual information with a single paradigm that uses natural videos as stimuli [7, 27-29, 35]. The studies have shown the unique value of using DNN and video-fMRI to map visual-field representations [7], the functional hierarchy of the visual cortex [7], cortical representations of categories [7, 27] or mid-level attributes (e.g. shapes or body parts) [27], and the hierarchical distribution of process memory [29]. As such, it provides an all-in-one strategy for mapping and characterizing various functional and computational aspects of vision. Although this strategy initially requires hours of video-fMRI data from each individual subject, we have recently developed a Bayesian transferring method to yield comparable results with only tens of minutes of video-fMRI data [28], making it practical for applications to many subjects in group studies. Through open source and data sharing, this dissertation also delivers a public resource to artificial intelligence and neuroscience communities, to promote positive, sustainable, and productive synergy between these two fields [7].

The DNN models used in previous studies [7-10, 22, 27, 30, 155] are all feedforward only. However, the brain contains both feedforward and feedback pathways, and their complex interactions give rise to visual perception, attention, and action [13, 220]. The interplay between feedforward and feedback connections is described by the predictive coding theory [14, 16, 204]. That is, the feedback connections from a higher visual area to a lower visual area carry predictions of lower-level neural activities, whereas the feedforward connections carry the residual errors between the predictions and the actual lower-level activities [16]. Such computations are supported by rich empirical evidence and the canonical microcircuitry of cortical columns [13]. Inspired by the theory of predictive coding, I proposed a PCN that combines feedforward, feedback and recurrent connections into a bi-directional and hierarchical network, which learns better representations for object recognition [39]. Therefore, the PCN sheds light on modeling the bi-directional feedforward and feedback processes for learning visual representations.

For future studies, relative to alternative models (CNN, RNN, VAE), the PCN offers a more comprehensive framework for modeling and mapping fMRI responses to natural videos. It will allow us to map the cortical hierarchies of spatial [7, 27, 28, 35] and temporal [29] representations, parsing the visual cortex into sub-areas or networks engaged in different levels of spatial or temporal processing. It will further allow us to separate and map feedforward and feedback pathways, and characterize their distinctive roles in natural vision. Although the focus on this project is on vision, the central idea is also applicable to other sensory systems. Natural hearing, speech and language processing are readily attainable goals [87, 221, 222].

Reverse engineering the brain in action is a common objective for neuroscience and artificial intelligence (AI) [1-3]. Understanding the brain will help guide and advance the development of next-generation AI. It will lead to detailed knowledge about the organization and connectivity of the human visual cortex to inform the design for deep learning. This dissertation proposed a strategy to compare brain-inspired AI against the brain itself [7, 27-29, 35, 39]. Notably, identifying the most effective rules for learning models is not only essential to machine learning, but also fundamental to human learning [223], which concerns how the human brain organizes information and learns new concepts from experiences.

REFERENCES

[1] D.D. Cox, T. Dean, Neural networks and neuroscience-inspired computer vision, Current Biology, 24 (2014) R921-R929.

[2] D.L. Yamins, J.J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex, Nature neuroscience, 19 (2016) 356-365.

[3] N. Kriegeskorte, Deep neural networks: a new framework for modeling biological vision and brain information processing, Annual Review of Vision Science, 1 (2015) 417-446.

[4] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature, 521 (2015) 436-444.

[5] J.Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE, 2015, pp. 4694-4702.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1-9.

[7] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, Z. Liu, Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision, Cerebral Cortex, DOI 10.1093/cercor/bhx268(2017) 1-25.

[8] U. Güçlü, M.A. van Gerven, Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream, Journal of Neuroscience, 35 (2015) 10005-10014.

[9] D.L. Yamins, H. Hong, C.F. Cadieu, E.A. Solomon, D. Seibert, J.J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex, Proceedings of the National Academy of Sciences, 111 (2014) 8619-8624.

[10] S.-M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation, PLoS Comput Biol, 10 (2014) e1003915.

[11] B.B. Averbeck, P.E. Latham, A. Pouget, Neural correlations, population coding and computation, Nature reviews neuroscience, 7 (2006) 358.

[12] T. Naselaris, K.N. Kay, S. Nishimoto, J.L. Gallant, Encoding and decoding in fMRI, Neuroimage, 56 (2011) 400-410.

[13] A.M. Bastos, W.M. Usrey, R.A. Adams, G.R. Mangun, P. Fries, K.J. Friston, Canonical microcircuits for predictive coding, Neuron, 76 (2012) 695-711.

[14] K. Friston, S. Kiebel, Predictive coding under the free-energy principle, Philosophical Transactions of the Royal Society B: Biological Sciences, 364 (2009) 1211-1221.

[15] Y. Huang, R.P. Rao, Predictive coding, Wiley Interdisciplinary Reviews: Cognitive Science, 2 (2011) 580-593.

[16] R.P. Rao, D.H. Ballard, Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects, Nature neuroscience, 2 (1999) 79.

[17] Y. Kamitani, F. Tong, Decoding the visual and subjective contents of the human brain, Nature neuroscience, 8 (2005) 679.

[18] J.-D. Haynes, G. Rees, Neuroimaging: decoding mental states from brain activity in humans, Nature Reviews Neuroscience, 7 (2006) 523.

[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 2012, pp. 1097-1105.

[20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, DOI (2014).

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.
[22] R.M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence, Scientific reports, 6 (2016).
[23] K.A. Norman, S.M. Polyn, G.J. Detre, J.V. Haxby, Beyond mind-reading: multi-voxel pattern analysis of fMRI data, Trends in cognitive sciences, 10 (2006) 424-430.
[24] K.N. Kay, T. Naselaris, R.J. Prenger, J.L. Gallant, Identifying natural images from human brain activity, Nature, 452 (2008) 352-355.
[25] T. Naselaris, R.J. Prenger, K.N. Kay, M. Oliver, J.L. Gallant, Bayesian reconstruction of natural images from human brain activity, Neuron, 63 (2009) 902-915.
[26] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, J.L. Gallant, Reconstructing visual experiences from brain activity evoked by natural movies, Current Biology, 21 (2011) 1641-1646.
[27] H. Wen, J. Shi, W. Chen, Z. Liu, Deep Residual Network Predicts Cortical Representation and Organization of Visual Features for Rapid Categorization, Scientific reports, 8 (2018) 3752.
[28] H. Wen, J. Shi, W. Chen, Z. Liu, Transferring and Generalizing Deep-Learning-based Neural Encoding Models across Subjects, bioRxiv, DOI (2017) 171017.
[29] J. Shi, H. Wen, Y. Zhang, K. Han, Z. Liu, Deep Recurrent Neural Network Reveals a Hierarchy of Process Memory during Dynamic Natural Vision, Human brain mapping, DOI https://doi.org/10.1002/hbm.24006(2018).
[30] M. Eickenberg, A. Gramfort, G. Varoquaux, B. Thirion, Seeing it all: Convolutional network layers map the function of the human visual system, NeuroImage, 152 (2017) 184-194.
[31] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, BMVC, 2015, pp. 6.
[32] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.
[33] P.O. Pinheiro, T.-Y. Lin, R. Collobert, P. Dollár, Learning to refine object segments, European Conference on Computer Vision, Springer, 2016, pp. 75-91.
[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248-255.
[35] K. Han, H. Wen, J. Shi, K.-H. Lu, Y. Zhang, Z. Liu, Variational autoencoder: An unsupervised model for modeling and decoding fMRI activity in visual cortex, bioRxiv, DOI (2017) 214247.
[36] D. George, J. Hawkins, Towards a mathematical theory of cortical micro-circuits, PLoS computational biology, 5 (2009) e1000532.
[37] A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science, Behavioral and brain sciences, 36 (2013) 181-204.
[38] J. Hohwy, The predictive mind, Oxford University Press2013.
[39] H. Wen, K. Han, J. Shi, Y. Zhang, E. Culurciello, Z. Liu, Deep Predictive Coding Network for Object Recognition, arXiv preprint arXiv:1802.04762, DOI (2018).
[40] C. Wacongne, J.-P. Changeux, S. Dehaene, A neuronal model of predictive coding accounting for the mismatch negativity, Journal of Neuroscience, 32 (2012) 3665-3678.
[41] L. de-Wit, B. Machilsen, T. Putzeys, Predictive coding and the neural response to predictable stimuli, Journal of Neuroscience, 30 (2010) 8702-8703.

[42] M.V. Srinivasan, S.B. Laughlin, A. Dubs, Predictive coding: a fresh view of inhibition in the retina, Proc. R. Soc. Lond. B, 216 (1982) 427-459.
[43] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, European conference on computer vision, Springer, 2014, pp. 818-833.
[44] B.E. Russ, D.A. Leopold, Functional MRI mapping of dynamic visual features during natural viewing in the macaque, Neuroimage, 109 (2015) 84-94.
[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, International Journal of Computer Vision, 115 (2015) 211-252.
[46] J.J. DiCarlo, D. Zoccolan, N.C. Rust, How does the brain solve visual object recognition?, Neuron, 73 (2012) 415-434.
[47] E.M. Callaway, Feedforward, feedback and inhibitory connections in primate visual cortex, Neural Networks, 17 (2004) 625-632.
[48] P.-O. Polack, D. Contreras, Long-range parallel processing and local recurrent activity in the visual cortex of the mouse, Journal of Neuroscience, 32 (2012) 11120-11131.
[49] M.F. Glasser, S.N. Sotiropoulos, J.A. Wilson, T.S. Coalson, B. Fischl, J.L. Andersson, J. Xu, S. Jbabdi, M. Webster, J.R. Polimeni, The minimal preprocessing pipelines for the Human Connectome Project, Neuroimage, 80 (2013) 105-124.
[50] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2014, pp. 806-813.
[51] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, R. Malach, Intersubject synchronization of cortical activity during natural vision, science, 303 (2004) 1634-1640.
[52] K.-H. Lu, S.-C. Hung, H. Wen, L. Marussich, Z. Liu, Influences of high-level features, gaze, and scene transitions on the reliability of BOLD responses to natural movie stimuli, PloS one, 11 (2016) e0161797.
[53] U. Güçlü, M.A. van Gerven, Increasingly complex representations of natural movies across the dorsal stream are shared between subjects, NeuroImage, DOI (2015).
[54] R.O. Abdollahi, H. Kolster, M.F. Glasser, E.C. Robinson, T.S. Coalson, D. Dierker, M. Jenkinson, D.C. Van Essen, G.A. Orban, Correspondences between retinotopic areas and myelin maps in human visual cortex, Neuroimage, 99 (2014) 509-524.
[55] J. Kubilius, S. Bracci, H.P.O. de Beeck, Deep neural networks as a computational model for human shape sensitivity, PLoS computational biology, 12 (2016) e1004896.
[56] T. Horikawa, Y. Kamitani, Generic decoding of seen and imagined objects using hierarchical visual features, Nature communications, 8 (2017).
[57] N. Kanwisher, J. McDermott, M.M. Chun, The fusiform face area: a module in human extrastriate cortex specialized for face perception, Journal of neuroscience, 17 (1997) 4302-4311.
[58] M.H. Johnson, Subcortical face processing, Nature Reviews Neuroscience, 6 (2005) 766.
[59] D. Adolf, S. Weston, S. Baecke, M. Luchtmann, J. Bernarding, S. Kropf, Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method, Frontiers in neuroinformatics, 8 (2014) 72.
[60] A.G. Huth, S. Nishimoto, A.T. Vu, J.L. Gallant, A continuous semantic space describes the representation of thousands of object and action categories across the human brain, Neuron, 76 (2012) 1210-1224.

[61] A.G. Huth, T. Lee, S. Nishimoto, N.Y. Bilenko, A.T. Vu, J.L. Gallant, Decoding the semantic content of natural movies from human brain activity, Frontiers in systems neuroscience, 10 (2016).
[62] M.F. Glasser, T.S. Coalson, E.C. Robinson, C.D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C.F. Beckmann, M. Jenkinson, A multi-modal parcellation of human cerebral cortex, Nature, DOI (2016).
[63] M.C.-K. Wu, S.V. David, J.L. Gallant, Complete functional characterization of sensory neurons by system identification, Annu. Rev. Neurosci., 29 (2006) 477-505.
[64] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene cnns, arXiv preprint arXiv:1412.6856, DOI (2014).
[65] D. Li, Visualization of deep convolutional neural networks, DOI (2016).
[66] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. MÃžller, How to explain individual classification decisions, Journal of Machine Learning Research, 11 (2010) 1803-1831.
[67] K. Hansen, D. Baehrens, T. Schroeter, M. Rupp, K.R. Müller, Visual Interpretation of Kernel-based prediction models, Molecular Informatics, 30 (2011) 817-826.
[68] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806, DOI (2014).
[69] J. Matyas, Random optimization, Automation and Remote control, 26 (1965) 246-253.
[70] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: A strategy employed by V1?, Vision research, 37 (1997) 3311-3325.
[71] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research, 15 (2014) 1929-1958.
[72] B.A. Wandell, S.O. Dumoulin, A.A. Brewer, Visual field maps in human cortex, Neuron, 56 (2007) 366-383.
[73] M. Bernstein, G. Yovel, Two neural pathways of face processing: a critical evaluation of current models, Neuroscience & Biobehavioral Reviews, 55 (2015) 536-546.
[74] B. Rossion, B. Hanseeuw, L. Dricot, Defining face perception areas in the human brain: a large-scale factorial fMRI face localizer analysis, Brain and cognition, 79 (2012) 138-157.
[75] E. Freud, D.C. Plaut, M. Behrmann, 'What'Is Happening in the Dorsal Visual Pathway, Trends in Cognitive Sciences, 20 (2016) 773-784.
[76] E.H. de Haan, A. Cowey, On the usefulness of 'what'and 'where'pathways in vision, Trends in cognitive sciences, 15 (2011) 460-466.
[77] M. Mur, D.A. Ruff, J. Bodurka, P. De Weerd, P.A. Bandettini, N. Kriegeskorte, Categorical, yet graded–single-image activation profiles of human category-selective cortical regions, Journal of Neuroscience, 32 (2012) 8649-8662.
[78] N. Kriegeskorte, M. Mur, D.A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, P.A. Bandettini, Matching categorical object representations in inferior temporal cortex of man and monkey, Neuron, 60 (2008) 1126-1141.
[79] T. Naselaris, D.E. Stansbury, J.L. Gallant, Cortical representation of animate and inanimate objects in complex natural scenes, Journal of Physiology-Paris, 106 (2012) 239-249.
[80] J.V. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, P. Pietrini, Distributed and overlapping representations of faces and objects in ventral temporal cortex, Science, 293 (2001) 2425-2430.

[81] T.A. Carlson, P. Schrater, S. He, Patterns of activity in the categorical representations of objects, Journal of cognitive neuroscience, 15 (2003) 704-717.
[82] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, S. Dehaene, Inverse retinotopy: inferring the visual content of images from brain activation patterns, Neuroimage, 33 (2006) 1104-1116.
[83] L. Itti, C. Koch, Computational modelling of visual attention, Nature reviews neuroscience, 2 (2001) 194.
[84] R. Desimone, J. Duncan, Neural mechanisms of selective visual attention, Annual review of neuroscience, 18 (1995) 193-222.
[85] Y. Miyawaki, H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H.C. Tanabe, N. Sadato, Y. Kamitani, Visual image reconstruction from human brain activity using a combination of multiscale local image decoders, Neuron, 60 (2008) 915-929.
[86] A. Hyvärinen, J. Hurri, P.O. Hoyer, Natural image statistics: a probabilistic approach to early computational vision, Springer2009.
[87] A.G. Huth, W.A. de Heer, T.L. Griffiths, F.E. Theunissen, J.L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex, Nature, 532 (2016) 453-458.
[88] J.G. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, JOSA A, 2 (1985) 1160-1169.
[89] T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization, Proceedings of the national academy of sciences, 104 (2007) 6424-6429.
[90] M.A. Goodale, A.D. Milner, Separate visual pathways for perception and action, Trends in neurosciences, 15 (1992) 20-25.
[91] T. Schenk, R.D. McIntosh, Do we have independent visual streams for perception and action?, DOI (2010).
[92] H. Kafaligonul, B.G. Breitmeyer, H. Öğmen, Feedforward and feedback processes in vision, Frontiers in psychology, 6 (2015) 279.
[93] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, Computer Vision (ICCV), 2015 IEEE International Conference on, IEEE, 2015, pp. 4489-4497.
[94] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, DOI (2015).
[95] A. Canziani, E. Culurciello, Visual attention with deep neural networks, Information sciences and systems (CISS), 2015 49th annual conference on, IEEE, 2015, pp. 1-3.
[96] J.J. DiCarlo, D.D. Cox, Untangling invariant object recognition, Trends in cognitive sciences, 11 (2007) 333-341.
[97] S. Thorpe, D. Fize, C. Marlot, Speed of processing in the human visual system, nature, 381 (1996) 520.
[98] D.C. Van Essen, C.H. Anderson, D.J. Felleman, Information processing in the primate visual system: an integrated systems perspective, Science, 255 (1992) 419.
[99] K. Grill-Spector, K.S. Weiner, The functional architecture of the ventral temporal cortex and its role in categorization, Nature Reviews Neuroscience, 15 (2014) 536-548.
[100] J.V. Haxby, J.S. Guntupalli, A.C. Connolly, Y.O. Halchenko, B.R. Conroy, M.I. Gobbini, M. Hanke, P.J. Ramadge, A common, high-dimensional model of the representational space in human ventral temporal cortex, Neuron, 72 (2011) 404-416.
[101] R.Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, I. Fried, Invariant visual representation by single neurons in the human brain, Nature, 435 (2005) 1102-1107.

[102] L.L. Chao, A. Martin, Representation of manipulable man-made objects in the dorsal stream, Neuroimage, 12 (2000) 478-484.
[103] S. Bracci, H.O. de Beeck, Dissociations and associations between shape and category representations in the two visual pathways, Journal of Neuroscience, 36 (2016) 432-444.
[104] V. Gallese, G. Lakoff, The brain's concepts: The role of the sensory-motor system in conceptual knowledge, Cognitive neuropsychology, 22 (2005) 455-479.
[105] A. Martin, The representation of object concepts in the brain, Annu. Rev. Psychol., 58 (2007) 25-45.
[106] A. Martin, GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain, Psychonomic bulletin & review, 23 (2016) 979-990.
[107] L.L. Chao, J.V. Haxby, A. Martin, Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects, Nature neuroscience, 2 (1999) 913-919.
[108] A.H. Bell, N.J. Malecek, E.L. Morin, F. Hadj-Bouziane, R.B. Tootell, L.G. Ungerleider, Relationship between functional magnetic resonance imaging-identified regions and neuronal category selectivity, Journal of Neuroscience, 31 (2011) 12229-12240.
[109] M. Brants, A. Baeck, J. Wagemans, H.P.O. de Beeck, Multiple scales of organization for object selectivity in ventral visual cortex, Neuroimage, 56 (2011) 1372-1381.
[110] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nature neuroscience, 2 (1999) 1019-1025.
[111] R. Kiani, H. Esteky, K. Mirpour, K. Tanaka, Object category structure in response patterns of neuronal population in monkey inferior temporal cortex, Journal of neurophysiology, 97 (2007) 4296-4309.
[112] B.Z. Mahon, S. Anzellotti, J. Schwarzbach, M. Zampini, A. Caramazza, Category-specific organization in the human brain does not require visual experience, Neuron, 63 (2009) 397-405.
[113] Z. Kourtzi, C.E. Connor, Neural representations for object perception: structure, category, and adaptive coding, Annual review of neuroscience, 34 (2011) 45-67.
[114] R. Epstein, N. Kanwisher, A cortical representation of the local visual environment, Nature, 392 (1998) 598-601.
[115] M.V. Peelen, P.E. Downing, Selectivity for the human body in the fusiform gyrus, Journal of neurophysiology, 93 (2005) 603-608.
[116] B.J. Devereux, A. Clarke, A. Marouchos, L.K. Tyler, Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects, Journal of Neuroscience, 33 (2013) 18906-18916.
[117] T.A. Carlson, R.A. Simmons, N. Kriegeskorte, L.R. Slevc, The emergence of semantic meaning in the ventral temporal pathway, Emergence, 26 (2013).
[118] A. Clarke, L.K. Tyler, Object-specific semantic coding in human perirhinal cortex, Journal of Neuroscience, 34 (2014) 4766-4775.
[119] A. Clarke, L.K. Tyler, Understanding what we see: how we derive meaning from vision, Trends in cognitive sciences, 19 (2015) 677-687.
[120] C.F. Cadieu, H. Hong, D.L. Yamins, N. Pinto, D. Ardila, E.A. Solomon, N.J. Majaj, J.J. DiCarlo, Deep neural networks rival the representation of primate IT cortex for core visual object recognition, PLoS Comput Biol, 10 (2014) e1003963.
[121] K.N. Kay, J. Winawer, A. Mezer, B.A. Wandell, Compressive spatial summation in human visual cortex, Journal of neurophysiology, 110 (2013) 481-494.

[122] C.J. Fox, G. Iaria, J.J. Barton, Defining the face processing network: optimization of the functional localizer in fMRI, Human brain mapping, 30 (2009) 1637-1651.
[123] S.-R. Afraz, R. Kiani, H. Esteky, Microstimulation of inferotemporal cortex influences face categorization, Nature, 442 (2006) 692-695.
[124] J. MacQueen, Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA., 1967, pp. 281-297.
[125] S. Gómez, P. Jensen, A. Arenas, Analysis of community structure in networks of correlated data, Physical Review E, 80 (2009) 016114.
[126] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, WordNet: An electronic lexical database, 49 (1998) 265-283.
[127] C. Fellbaum, WordNet, Wiley Online Library1998.
[128] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, 2013, pp. 3111-3119.
[129] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532-1543.
[130] N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis–connecting the branches of systems neuroscience, Frontiers in systems neuroscience, 2 (2008).
[131] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579, DOI (2015).
[132] S.M. Smith, C.F. Beckmann, J. Andersson, E.J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D.A. Feinberg, L. Griffanti, M.P. Harms, Resting-state fMRI in the human connectome project, Neuroimage, 80 (2013) 144-168.
[133] S.-M. Khaligh-Razavi, L. Henriksson, K. Kay, N. Kriegeskorte, Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models, Journal of Mathematical Psychology, 76 (2017) 184-197.
[134] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1717-1724.
[135] B.Z. Mahon, A. Caramazza, What drives the organization of object knowledge in the brain?, Trends in cognitive sciences, 15 (2011) 97-103.
[136] A. Caramazza, J.R. Shelton, Domain-specific knowledge systems in the brain: The animate-inanimate distinction, Journal of cognitive neuroscience, 10 (1998) 1-34.
[137] T.J. Andrews, D.M. Watson, G.E. Rice, T. Hartley, Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway, Journal of Vision, 15 (2015) 3-3.
[138] D.D. Coggan, W. Liu, D.H. Baker, T.J. Andrews, Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information, Neuroimage, 135 (2016) 107-114.
[139] D.M. Watson, M. Hymers, T. Hartley, T.J. Andrews, Patterns of neural response in scene-selective regions of the human brain are affected by low-level manipulations of spatial frequency, NeuroImage, 124 (2016) 107-117.

[140] D. Proklova, D. Kaiser, M.V. Peelen, Disentangling representations of object shape and object category in human visual cortex: The animate–inanimate distinction, Journal of cognitive neuroscience, DOI (2016).
[141] D. Kaiser, D.C. Azzalini, M.V. Peelen, Shape-independent object category responses revealed by MEG and fMRI decoding, Journal of neurophysiology, 115 (2016) 2246-2250.
[142] C. Baldassi, A. Alemi-Neissi, M. Pagan, J.J. DiCarlo, R. Zecchina, D. Zoccolan, Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons, PLoS computational biology, 9 (2013) e1003167.
[143] D.M. Drucker, G.K. Aguirre, Different spatial scales of shape similarity representation in lateral and ventral LOC, Cerebral Cortex, 19 (2009) 2269-2280.
[144] J. Haushofer, M.S. Livingstone, N. Kanwisher, Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity, PLoS biology, 6 (2008) e187.
[145] T. Konkle, A. Oliva, A real-world size organization of object responses in occipitotemporal cortex, Neuron, 74 (2012) 1114-1124.
[146] S. Gabay, E. Kalanthroff, A. Henik, N. Gronau, Conceptual size representation in ventral visual cortex, Neuropsychologia, 81 (2016) 198-206.
[147] J.W. Peirce, Understanding mid-level representations in visual processing, Journal of Vision, 15 (2015) 5-5.
[148] H.B. Barlow, Unsupervised learning, Neural computation, 1 (1989) 295-311.
[149] S. Park, T.F. Brady, M.R. Greene, A. Oliva, Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes, Journal of Neuroscience, 31 (2011) 1333-1340.
[150] T. Trappenberg, Fundamentals of computational neuroscience, OUP Oxford2009.
[151] L. Paninski, J. Pillow, J. Lewi, Statistical models for neural encoding, decoding, and optimal stimulus design, Progress in brain research, 165 (2007) 493-507.
[152] M. Chen, J. Han, X. Hu, X. Jiang, L. Guo, T. Liu, Survey of encoding and decoding of visual stimulus via FMRI: an image analysis perspective, Brain imaging and behavior, 8 (2014) 7-23.
[153] T.C. Kietzmann, P. McClure, N. Kriegeskorte, Deep Neural Networks In Computational Neuroscience, bioRxiv, DOI (2017) 133504.
[154] M. van Gerven, Computational Foundations of Natural Intelligence, bioRxiv, DOI (2017) 166785.
[155] K. Seeliger, M. Fritsche, U. Güçlü, S. Schoenmakers, J.-M. Schoffelen, S. Bosch, M. van Gerven, Convolutional neural network-based encoding and decoding of visual object recognition in space and time, NeuroImage, DOI (2017).
[156] B.R. Conroy, B.D. Singer, J.S. Guntupalli, P.J. Ramadge, J.V. Haxby, Inter-subject alignment of human cortical anatomy using functional connectivity, NeuroImage, 81 (2013) 400-411.
[157] G. Raz, M. Svanera, N. Singer, G. Gilam, M.B. Cohen, T. Lin, R. Admon, T. Gonen, A. Thaler, R.Y. Granot, Robust inter-subject audiovisual decoding in functional magnetic resonance imaging using high-dimensional regression, Neuroimage, 163 (2017) 244-263.
[158] J.L. Teeters, K.D. Harris, K.J. Millman, B.A. Olshausen, F.T. Sommer, Data sharing for computational neuroscience, Neuroinformatics, 6 (2008) 47-55.
[159] D.C. Van Essen, S.M. Smith, D.M. Barch, T.E. Behrens, E. Yacoub, K. Ugurbil, W.-M.H. Consortium, The WU-Minn human connectome project: an overview, Neuroimage, 80 (2013) 62-79.

[160] D.N. Paltoo, L.L. Rodriguez, M. Feolo, E. Gillanders, E.M. Ramos, J. Rutter, S. Sherry, V.O. Wang, A. Bailey, R. Baker, Data use under the NIH GWAS data sharing policy and future directions, Nature genetics, 46 (2014) 934.
[161] R.A. Poldrack, K.J. Gorgolewski, Making big data open: data sharing in neuroimaging, Nature neuroscience, 17 (2014) 1510-1517.
[162] J. Fan, F. Han, H. Liu, Challenges of big data analysis, National science review, 1 (2014) 293-314.
[163] Ó. Fontenla-Romero, B. Guijarro-Berdiñas, D. Martinez-Rego, B. Pérez-Sánchez, D. Peteiro-Barral, Online machine learning, Efficiency and Scalability Methods for Computational Intellect, 27 (2013).
[164] J.S. Guntupalli, M. Hanke, Y.O. Halchenko, A.C. Connolly, P.J. Ramadge, J.V. Haxby, A model of representational spaces in human cortex, Cerebral cortex, 26 (2016) 2919-2934.
[165] H. Zha, H.D. Simon, On updating problems in latent semantic indexing, SIAM Journal on Scientific Computing, 21 (1999) 782-791.
[166] H. Zhao, P.C. Yuen, J.T. Kwok, A novel incremental principal component analysis and its application for face recognition, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36 (2006) 873-886.
[167] G. St-Yves, T. Naselaris, The feature-weighted receptive field: An interpretable encoding model for complex feature spaces, NeuroImage, DOI (2017).
[168] M. Sahani, J.F. Linden, Evidence optimization techniques for estimating stimulus-response functions, Advances in neural information processing systems, 2003, pp. 317-324.
[169] S. Geisser, Predictive inference, CRC press1993.
[170] P.H. Schönemann, A generalized solution of the orthogonal Procrustes problem, Psychometrika, 31 (1966) 1-10.
[171] F.M. Dias, A. Antunes, J. Vieira, A.M. Mota, Implementing the levenberg-marquardt algorithm on-line: A sliding window approach with early stopping, IFAC Proceedings Volumes, 37 (2004) 49-54.
[172] J. Chen, Y.C. Leong, C.J. Honey, C.H. Yong, K.A. Norman, U. Hasson, Shared memories reveal shared structure in neural activity across individuals, Nature neuroscience, 20 (2017) 115-125.
[173] J. Goense, Y. Bohraus, N.K. Logothetis, fMRI at high spatial resolution: implications for BOLD-models, Frontiers in computational neuroscience, 10 (2016).
[174] E. Yacoub, N. Harel, K. Uğurbil, High-field fMRI unveils orientation columns in humans, Proceedings of the National Academy of Sciences, 105 (2008) 10607-10612.
[175] M.R. Hodge, W. Horton, T. Brown, R. Herrick, T. Olsen, M.E. Hileman, M. McKay, K.A. Archie, E. Cler, M.P. Harms, ConnectomeDB—sharing human brain connectivity data, Neuroimage, 124 (2016) 1102-1107.
[176] J.O. Berger, Statistical decision theory and Bayesian analysis, Springer Science & Business Media2013.
[177] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, arXiv preprint arXiv:1709.01507, DOI (2017).
[178] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, arXiv preprint arXiv:1707.07012, DOI (2017).
[179] S. Zagoruyko, N. Komodakis, Wide residual networks, arXiv preprint arXiv:1605.07146, DOI (2016).

[180] D.H. Hubel, T.N. Wiesel, Receptive fields and functional architecture of monkey striate cortex, The Journal of physiology, 195 (1968) 215-243.
[181] D.J. Felleman, D.E. Van, Distributed hierarchical processing in the primate cerebral cortex, Cerebral cortex (New York, NY: 1991), 1 (1991) 1-47.
[182] O. Sporns, J.D. Zwi, The small world of the cerebral cortex, Neuroinformatics, 2 (2004) 145-162.
[183] N.K. Logothetis, D.L. Sheinberg, Visual object recognition, Annual review of neuroscience, 19 (1996) 577-621.
[184] D. Wyatte, D.J. Jilk, R.C. O'Reilly, Early recurrent feedback facilitates visual object recognition under challenging conditions, Frontiers in psychology, 5 (2014) 674.
[185] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, DOI (2009).
[186] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, NIPS workshop on deep learning and unsupervised feature learning, 2011, pp. 5.
[187] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86 (1998) 2278-2324.
[188] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural computation, 29 (2017) 2352-2449.
[189] R.C. O'Reilly, D. Wyatte, S. Herd, B. Mingus, D.J. Jilk, Recurrent processing during object recognition, Frontiers in psychology, 4 (2013) 124.
[190] C.J. Spoerer, P. McClure, N. Kriegeskorte, Recurrent convolutional neural networks: a better model of biological object recognition, Frontiers in psychology, 8 (2017) 1551.
[191] M. Liang, X. Hu, Recurrent convolutional neural network for object recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3367-3375.
[192] M.F. Stollenga, J. Masci, F. Gomez, J. Schmidhuber, Deep networks with internal selective attention through feedback connections, Advances in Neural Information Processing Systems, 2014, pp. 3545-3553.
[193] A. Canziani, E. Culurciello, Cortexnet: a generic network family for robust visual temporal representations, arXiv preprint arXiv:1706.02735, DOI (2017).
[194] C. Gómez, J.T. Lizier, M. Schaum, P. Wollstadt, C. Grützner, P. Uhlhaas, C.M. Freitag, S. Schlitt, S. Bölte, R. Hornero, Reduced predictable information in brain signals in autism spectrum disorder, Frontiers in neuroinformatics, 8 (2014) 9.
[195] A.M. Bastos, J. Vezoli, C.A. Bosman, J.-M. Schoffelen, R. Oostenveld, J.R. Dowdall, P. De Weerd, H. Kennedy, P. Fries, Visual areas exert feedforward and feedback influences through distinct frequency channels, Neuron, 85 (2015) 390-401.
[196] G. Michalareas, J. Vezoli, S. Van Pelt, J.-M. Schoffelen, H. Kennedy, P. Fries, Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas, Neuron, 89 (2016) 384-397.
[197] W. Sedley, P.E. Gander, S. Kumar, C.K. Kovach, H. Oya, H. Kawasaki, M.A. Howard III, T.D. Griffiths, Neural signatures of perceptual inference, Elife, 5 (2016).
[198] S. van Pelt, L. Heil, J. Kwisthout, S. Ondobaka, I. van Rooij, H. Bekkering, Beta-and gamma-band activity reflect predictive coding in the processing of causal events, Social cognitive and affective neuroscience, 11 (2016) 973-980.
[199] R.P. Rao, D.H. Ballard, Dynamic model of visual recognition predicts neural response properties in the visual cortex, Neural computation, 9 (1997) 721-763.

[200] K. Friston, Hierarchical models in the brain, PLoS computational biology, 4 (2008) e1000211.
[201] R. Chalasani, J.C. Principe, Deep predictive coding networks, arXiv preprint arXiv:1301.3541, DOI (2013).
[202] W. Lotter, G. Kreiman, D. Cox, Deep predictive coding networks for video prediction and unsupervised learning, arXiv preprint arXiv:1605.08104, DOI (2016).
[203] M.W. Spratling, A hierarchical predictive coding model of object recognition in natural images, Cognitive computation, 9 (2017) 151-167.
[204] M.W. Spratling, Predictive coding as a model of biased competition in visual attention, Vision research, 48 (2008) 1391-1408.
[205] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807-814.
[206] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, International conference on machine learning, 2015, pp. 448-456.
[207] V. Dumoulin, F. Visin, A guide to convolution arithmetic for deep learning, arXiv preprint arXiv:1603.07285, DOI (2016).
[208] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, DOI (2017).
[209] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, European Conference on Computer Vision, Springer, 2016, pp. 630-645.
[210] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, arXiv preprint arXiv:1302.4389, DOI (2013).
[211] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, DOI (2014).
[212] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint arXiv:1312.4400, DOI (2013).
[213] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, Artificial Intelligence and Statistics, 2015, pp. 562-570.
[214] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, arXiv preprint arXiv:1412.6550, DOI (2014).
[215] R.K. Srivastava, K. Greff, J. Schmidhuber, Training very deep networks, Advances in neural information processing systems, 2015, pp. 2377-2385.
[216] G. Huang, Z. Liu, K.Q. Weinberger, L. van der Maaten, Densely connected convolutional networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3.
[217] M.D. Zeiler, R. Fergus, Stochastic pooling for regularization of deep convolutional neural networks, arXiv preprint arXiv:1301.3557, DOI (2013).
[218] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, R. Fergus, Regularization of neural networks using dropconnect, International Conference on Machine Learning, 2013, pp. 1058-1066.
[219] B.R. Conway, D.Y. Tsao, Color architecture in alert macaque cortex revealed by FMRI, Cerebral Cortex, 16 (2005) 1604-1613.
[220] K. Friston, The free-energy principle: a unified brain theory?, Nature Reviews Neuroscience, 11 (2010) 127.
[221] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, Deep neural networks for acoustic modeling in speech recognition:

The shared views of four research groups, IEEE Signal Processing Magazine, 29 (2012) 82-97.
[222] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, Deep speech 2: End-to-end speech recognition in english and mandarin, International Conference on Machine Learning, 2016, pp. 173-182.
[223] N. Kraus, T. White-Schwoch, Unraveling the biology of auditory learning: a cognitive–sensorimotor–reward framework, Trends in Cognitive Sciences, 19 (2015) 642-654.

VITA

Haiguang Wen is a PhD candidate at the School of Electrical and Computer Engineering, Purdue University, IN, USA. He received his B.S. degree in Information Science and Communication Engineering from Zhejiang University, Hangzhou, Zhejiang, China, in 2013. He joined the Laboratory of Integrated Brain Imaging, Purdue, in August 2013.

His research interests are in signal and image processing, computational neuroscience, and deep learning. During 2013 to 2014, his work was on the fractal (or scale-free) dynamics in the brain activity. After that, he started working on the computational neuroscience and deep learning. He received the Merit Award at the international conference, the Organization of Human Brain Mapping (OHBM), in 2016 and 2018, and the Neuroscience Research Travel Award from Purdue Institute for Integrative in 2016.

PUBLICATIONS

Journal Articles

[1] **Wen H.,** Shi J., Zhang Y., Lu K-H, Liu Z. (2017). "Neural encoding and decoding with deep learning for dynamic natural vision." *Cerebral Cortex*, doi:10.1093/cercor/bhx268.

[2] **Wen H.**, Shi J., Chen W., Liu Z. (2017). "Deep residual network predicts cortical representation and organization of visual features for rapid categorization." *Scientific Reports*, *8*(1), 3752.

[3] **Wen H.**, Shi J., Chen W., Liu Z. (2017). "Transferring and generalizing deep-learning-based neural encoding models across subjects." doi: 10.1101/171017. [revision under review by *NeuroImage*]

[4] **Wen H.**, Han K., Shi J., Zhang Y., Culurciello E., Liu Z. (2018). "Deep predictive coding network for object recognition," arXiv:1802.04762. [under review by *International Conference on Machine Learning, ICML*]

[5] **Wen H.**, Liu Z. (2016). "Broadband electrophysiological dynamics contribute to global resting-state fMRI signal." *Journal of Neuroscience*, 36(22): 6030-6040.

[6] **WenH.**, LiuZ. (2016). "Separating fractal and oscillatory components in the power spectrum of neurophysiological signal." *Brain Topography*, 29(1): 13-26.

[7] Shi J., **Wen H.**, Zhang Y., Han K., Liu Z. (2018). "Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision." *Human Brain Mapping*, doi: 10.1002/hbm.24006.

[8] Han K., **Wen H.**, Shi J., Lu K-H, Zhang Y., Liu Z. (2017). "Variational Autoencoder: An Unsupervised Model for Modeling and Decoding fMRI Activity in Visual Cortex." doi:10.1101/214247. [under review by *NeuroImage*]

[9] Zhang Y., Chen G., **Wen H.**, Lu K-H., Liu Z. (2017). "Musical Imagery Involves the Wernicke's Area in Bilateral and Anti-Correlated Network Interactions in Musicians." *Scientific Reports*, 7: 17066.

[10] Marussich L., Lu K-H, **Wen H.**, Liu Z. (2017). "Mapping white-matter functional organization at rest and during naturalistic visual perception." *NeuroImage*, 146: 1128-1141.

[11] Lu K-H, Jeong J-H, **Wen H.**, Liu Z. (2017). "Spontaneous activity in the visual cortex is organized by visual streams." *Human Brain Mapping*, 38(9): 4613-4630.

[12] Lu K-H, Hung S., **Wen H.**, Marussich L., Liu Z. (2016). "Influences of high-level features, gaze, and scene transitions on the reliability of BOLD responses to natural movie stimuli." *PLoS ONE*, 11(8): e0161797.

[13] Lynch L., Lu K-H., **Wen H**., Zhang Y, Saykin AJ., Liu Z. (2018). "Task-evoked functional connectivity does not explain functional connectivity differences between rest and task conditions," doi: 10.1101/252759. [under review by *Human Brain Mapping*]

Conference Proceedings

[1] **Wen H.**, Shi J., Wei C., Liu Z., "Learning Transferable and Generalizable Neural Encoding Models for Natural Vision." In Proceedings of the 24rd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2018. **[Merit Abstract Award, Oral and Poster]**

[2] **Wen H.**, Shi J., Zhang Y., Han K., Liu Z., "Distributed cortical networks represent visual object categories." In Proceedings of the 23rd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2017. **[Poster]**

[3] **Wen H.**, Shi J., Lu K-H, Zhang Y., Marussich L., Liu Z., "Decode cortical fMRI activity to reconstruct naturalistic movie via deep learning." In Proceedings of the 22nd Annual Meeting of the OHBM, 2016. **[Merit Abstract Award, Oral and Poster]**

[4] **Wen H.**, Jeong JY, Liu Z., "Intrinsic functional networks within visual cortex supports naturalistic visual perception." In Proceedings of the 22nd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2016. [**Neuroscience Research Travel Award, Poster**]

[5] **Wen H.**, Liu Z., "Distinct neurophysiological correlates of global vs. local resting state fMRI networks." Intl. Soc. Magn. Reson. Med. Annual Scientific Meeting, 2015. **[Power-Pitch Highlight, Oral and Poster]**

[6] **Wen H.**, Liu Z., "Functional networks observed with scale-free and oscillatory cortical activity." Resting State Brain Connectivity Conference, 2014. **[Poster]**

[7] Shi J., **Wen H.**, Zhang Y., Han K., Liu Z., "Deep recurrent neural network reveals a hierarchy of temporal receptive windows in the visual cortex." In Proceedings of the 23rd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2017. **[Merit Abstract Award, Oral and Poster]**

[8] Zhang Y., Kim J-H, **Wen H.**, Liu Z. "High Gamma Electrocorticography in Superior Temporal Gyrus Represents Words during Natural Speech. In Proceedings of the 24nd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2018. **[Merit Abstract Award, Oral and Poster]**

[9] Lu K-H, Jeong JY, **Wen H.**, Liu Z., "Spontaneous activity patterns reveal non-retinotopic functional parcellation and organization of human visual cortex." the Scientific Meeting of the International Society of Magnetic Resonance for Medicine (ISMRM), 2017. **[Summa Cum Laude Award, Oral and Poster]**

[10] Zhang Y., **Wen H.**, Lu K-H, Liu Z., " Common and Distinct Cortical Network Bases of Musical Perception and Imagery." In Proceedings of the 23rd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2017. **[Poster]**

[11] Han K., **Wen H.**, Shi J., Lu K-H, Liu Z., "Decoding Cortical Activity with Variational Autoencoder Supports Direct Visual Reconstruction." In Proceedings of the 23rd Annual Meeting of the OHBM, 2017. **[Poster]**

[12] Kim J-H, **Wen H.**, Zhang Y., Liu Z., "Development of new EEG-fMRI source imaging method for continuous task paradigm." In Proceedings of the 23rd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2017. **[Poster]**

[13] Shi J., **Wen H.,** Lu K-H, Liu Z., "Mapping Neural Representation of Hierarchical Visual Features during Natural Movie Stimuli," In Proceedings of the 22nd Annual Meeting of the OHBM, 2016. **[Travel Award, Poster]**

[14] Lu K-H, **Wen H.**, Liu Z., "Sources of reliable fMRI responses to natural movie stimuli." In Proceedings of the 22nd Annual Meeting of the OHBM, Geneva, Switzerland, 2016. **[Poster]**

[15] Marussich L., Lu K-H, **Wen H.**, Liu Z., "Hierarchical clusters of white-matter fMRI are coupled with cortical visual networks." In Proceedings of the 22nd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2016. **[Poster]**

[16] Han K., **Wen H.**, Zhang Y., Liu Z. "Comparing Deep Neural Network Based Encoding Models for Predicting Movie induced Cortical Activities", In Proceedings of the 24nd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2018. **[Poster]**

[17] Lu K-H, **Wen H.**, Liu Z. "Fine-scale ICA Reveals Retinotopic Organization in the Visual Cortex under Natural Vision", In Proceedings of the 24nd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2018. **[Poster]**

[18] Lynch LK, Lu K-H, **Wen H.**, Zhang Y., Saykin AJ, Liu Z. "Task-Evoked FC Does Not Explain FC Differences Between Rest and Task Conditions", In Proceedings of the 24nd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2018. **[Poster]**

[19] Wang W., Han K., **Wen H.**, Liu Z. "A web-based platform for predicting brain responses based on deep neural networks", In Proceedings of the 24nd Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2018. **[Poster]**