How does the brain represent the semantic content of an image?

Huawei Xu[a,b], **Ming Liu**[a,b,c,1], **and Delong Zhang**[a,b,c,1]

[a]Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education, China 510631; [b]School of Psychology, Center for Studies of Psychological Application, and Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, China 510631; [c]Plateau Brain Science Research Center, South China Normal University/Tibet University, Guangzhou/Lhasa, China 510631

This manuscript was compiled on July 3, 2021

Using deep neural networks (DNNs) as models to explore the biological brain is controversial, which is mainly due to the impenetrability of DNNs. Inspired by neural style transfer, we circumvented this problem by using deep features that were given a clear meaning—the representation of the semantic content of an image. Using encoding models and the representational similarity analysis, we quantitatively showed that the deep features which represented the semantic content of an image mainly modulated the activity of voxels in the early visual areas (V1, V2, and V3) and these features were essentially depictive but also propositional. This result is in line with the core viewpoint of the grounded cognition to some extent, which suggested that the representation of information in our brain is essentially depictive and can implement symbolic functions naturally.

deep neural networks | neural style transfer | early visual areas | ground cognition

Deep neural networks (DNNs) for image recognition provided an important tool for understanding the nature of visual object recognition (1–5). This is not only because DNNs are currently the only models known to achieve near human-level performance in object recognition, but also because they have the properties such as the hierarchical organization and the parallel distributed processing which are similar to the visual ventral stream—key circuits that underlie visual object recognition (6, 7). Using DNNs as computational models, researchers found that DNNs could predict brain activity of visual processing across multiple hierarchical levels at unprecedented accuracy for both macaque (8–11) and human (12–15) wherein later layers in DNNs better predict higher areas of the visual ventral stream. The predictive power of DNNs made "mind-reading" possible (16–18) and promoted the integration of neuroscience and artificial intelligence (19, 20).

Besides the predictive power, an ideal model should also possess the explanatory power, which means that we should know how the model works (21). This is not the case of DNNs. DNNs are essentially black boxes and we can not understand how the input data were transformed into model output (22). This is mainly due to the end-to-end learning and the huge number of parameters in DNNs (the complex architectures of DNNs). For example, AlexNet has about 60 million self-learned parameters (23) and VGG16 has 138 million self-learned parameters (24). Even though we know the exact value of all parameters for each input, we still can not understand what do these parameters really mean. So using DNNs as models to explore the biological brain is something like replacing a black box with another, the lack of explanatory made it controversial (1). To open the black box and look inside, researchers developed methods such as network dissection (25) and visualization (26–30), and experimented with network architecture (31), learning algorithm (32, 33), and input statistics (34). But none of them can directly explain the meaning of the parameters (deep features) learned by DNNs.

However, an interesting and successful application of DNNs may give us a hint about the meanings of some deep features. Neural style transfer (NST) is a computer vision technique that allows us to render the semantic content of an image in the style of another (35–37). Using NST, for example, we can blend a photo with van Gogh's "Sunflowers" to get a new image which preserve the content of the photo but looks like if it was painted by van Gogh. According to the seminal work of (36), the implementation of the original NST algorithm was based on a DNN optimized for object recognition—VGG19. This process took two images, a content image and a style image. First, two images were fed into the pre-trained VGG19 model to extract feature maps, respectively. Second, the feature maps of the conv4_2 layer of the content image were selected as the semantic content representation. Third, the feature maps of the conv1_1 layer, conv2_1 layer, conv3_1 layer, conv4_1 layer, and conv5_1 layer of the style image were selected to compute the Gram matrix as the style representation. Last, through jointly minimizing the distance of the feature representations of a white noise image from the content representation and the style representation (feature inversion using the same VGG19 model), a new image was generated which simultaneously match the content of the content image and the style of the style image. The key to NST lies in the ability to extract representation from an image which

Significance Statement

Using deep neural networks (DNNs) as models to explore the biological brain is something like replacing a black box with another, the impenetrability of DNNs made it controversial. However, an interesting and successful application of DNNs—Neural Style Transfer—may help us to circumvent this problem, which can separate image content from style. In this study, we quantitatively showed that the deep features which represented the semantic content of images mainly modulated the activity of voxels in the early visual areas and these features were essentially depictive but also propositional. Our result provided preliminary evidence that information in our brain is essentially depictive and can implement symbolic functions naturally.

H.X., M.L., and D.Z. designed research; H.X. and D.Z. performed research; and H.X. and D.Z. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence should be addressed. E-mail: lim@scnu.edu.cn(ML); delong.zhang@m.scnu.edu.cn(DZ)

DRAFT

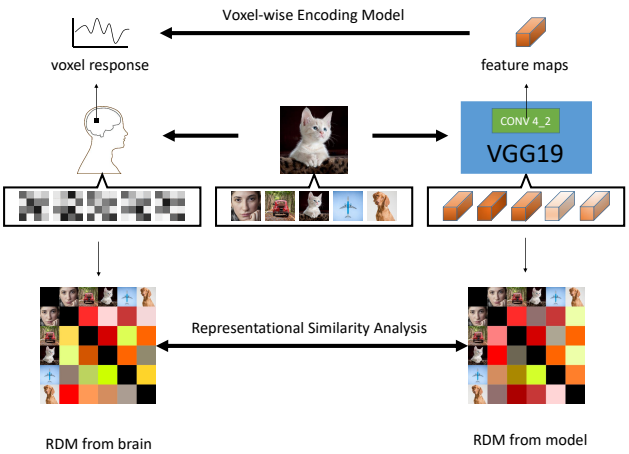

Fig. 1. The Experimental Workfolw: encoding models and representational similarity analysis.

explicitly separate image content from style (36).

In this study, we focused on the feature maps of the conv4_2 layer of the VGG19, which were selected as the representation of the semantic content of an image in the original NST algorithm. Although there was no clear explanation about why choose the layer conv4_2 as the semantic content representation of an image, the original NST algorithm was indeed effective which led to lots of follow-up studies and many successful industrial applications (e.g., Prisma). So it gave us an opportunity to explore the question of how does the brain represent the semantic content of an image. We used voxel-wise encoding models (38–40) to answer this question, which could test hypotheses about how information is represented in our brain. The results showed that, the deep features, which represented the semantic content of an image, mainly modulated the activity of voxels in the early visual areas (V1, V2, and V3). These semantics-related features mainly modulated the voxels in the early visual areas rather than those in the higher visual areas naturally led us to another question—what these features really are. For this question, we constructed encoding models based on Gabor features which also modulated the activity of voxels in the early visual areas (41) to compare it with encoding models using the deep features and used representational similarity analysis (RSA, 42–44) to explore the representational similarity between the representation of the semantic content of an image and other representations such as the representation of semantics and the representation of Gabor features. We found that these features were essentially depictive but also propositional. It is in line with the core viewpoint of the grounded cognition (45–47) to some extent, which suggested that the representation of information in our brain was essentially depictive and could implement symbolic functions naturally.

Results

The fMRI data we used was from (16), which contained a training set (subject viewed 1200 natural images), a testing set (subject viewed 50 natural images), and an imagery set (subject imagined objects according to 50 nouns) for each of 5 subjects. In addition, 7 ROIs (V1, V2, V3, V4, LOC, FFA, and PPA) were identified for each subject (The fMRI data only contained these ROIs). To explore how does the brain represent the semantic content of an image, we extracted deep features which represented the semantic content of images from the conv4_2 layer of a pre-trained VGG19 and constructed lasso-regularized linear models to predict voxel repsonses from these features using the training sets for each voxel in each subject. Once models were fitted, we used the testing sets and the imagery sets to evaluate models with Pearson's correlation coefficient (r) for each voxel and decoding performance (identifying stimuli from measured brain activity) for all survived voxels. After that we analysed survived models to find the relationship between features and voxels. Because the models showed that these deep features mainly modulated the activity of voxels in the early visual areas, we then compared it with encoding models using Gabor features—the low-level visual features which have been proved to tune simple cells in primary visual cortex (48) and modulate the activity of voxels in the early visual areas (41). To further explore what the representation of the semantic content of an image really is, we constructed three types of representational distance matrixs (RDMs)—the RDMs from VGG19 (the conv5_4 layer and the fc2 layer), the RDMs from brain activity (7 ROIs for each subject), and the RDMs directly from stimuli (Gabor features, silhouette, and semantics)—to compare them with the RDM of the conv4_2 layer using the testing sets and the imagery sets (Fig. 1).

The deep features which represented the semantic content of an image mainly modulated the activity of voxels in the early visual areas. Because only 4 models survived in the imagery sets (2 for S2, 1 for S3, and 1 for S5), the following analysis mainly focused on the testing sets. The number of survived models (voxels) in the testing sets was 201 of 4466 for S1, 356 of 4404 for S2, 789 of 4643 for S3, 701 of 4133 for S4, and 369 of 4370 for S5. The distribution of survived models in ROIs was different between 5 Subjects (Fig. 2.A for S1, the result of other subjects can be found in SI Appendix, same below in all figures). There were only 4 ROIs left in S1: V1 (115), V2 (104), V3 (36), and V4 (5). The number of ROIs left in S5 was 6: V1 (223), V2 (185), V3 (53), V4 (14), LOC (1), and PPA (7). All the 7 ROIs left in S2, S3, and S4. For S2, V1 remained 210 voxels; V2 remained 164 voxels; V3 remained 54 voxels; V4 remained 4 voxels; LOC remained 11 voxels; FFA remained 4 voxels; PPA remained 1 voxels. For S3, V1 remained 334 voxels; V2 remained 356 voxels; V3 remained 180 voxels; V4 remained 58 voxels; LOC remained 48 voxels; FFA remained 43 voxels; PPA remained 20 voxels. For S4, V1 remained 328 voxels; V2 remained 308 voxels; V3 remained 177 voxels; V4 remained 28 voxels; LOC remained 20 voxels; FFA remained 37 voxels; PPA remained 17 voxels (Because some voxels simultaneously belonged to two different ROIs, the sum of the number of all the voxels in different ROIs may be larger than the total number of the survived voxels for each subject).

Another measurement for prediction accuracy in different ROIs is the median of Pearson's correlation coefficients of all survived voxels in each ROI (Fig. 2.B for S1). The prediction accuracies of the V1, V2, V3, and V4 were 0.60, 0.58, 0.55, and 0.55 for S1; The prediction accuracies of the V1, V2, V3, V4, LOC, FFA, and PPA were 0.64, 0.58, 0.57, 0.52, 0.54, 0.52, and 0.50 for S2; The prediction accuracies of the V1, V2, V3, V4, LOC, FFA, and PPA were 0.68, 0.65, 0.63, 0.58, 0.57, 0.58, and 0.57 for S3; The prediction accuracies of the V1, V2,

DRAFT

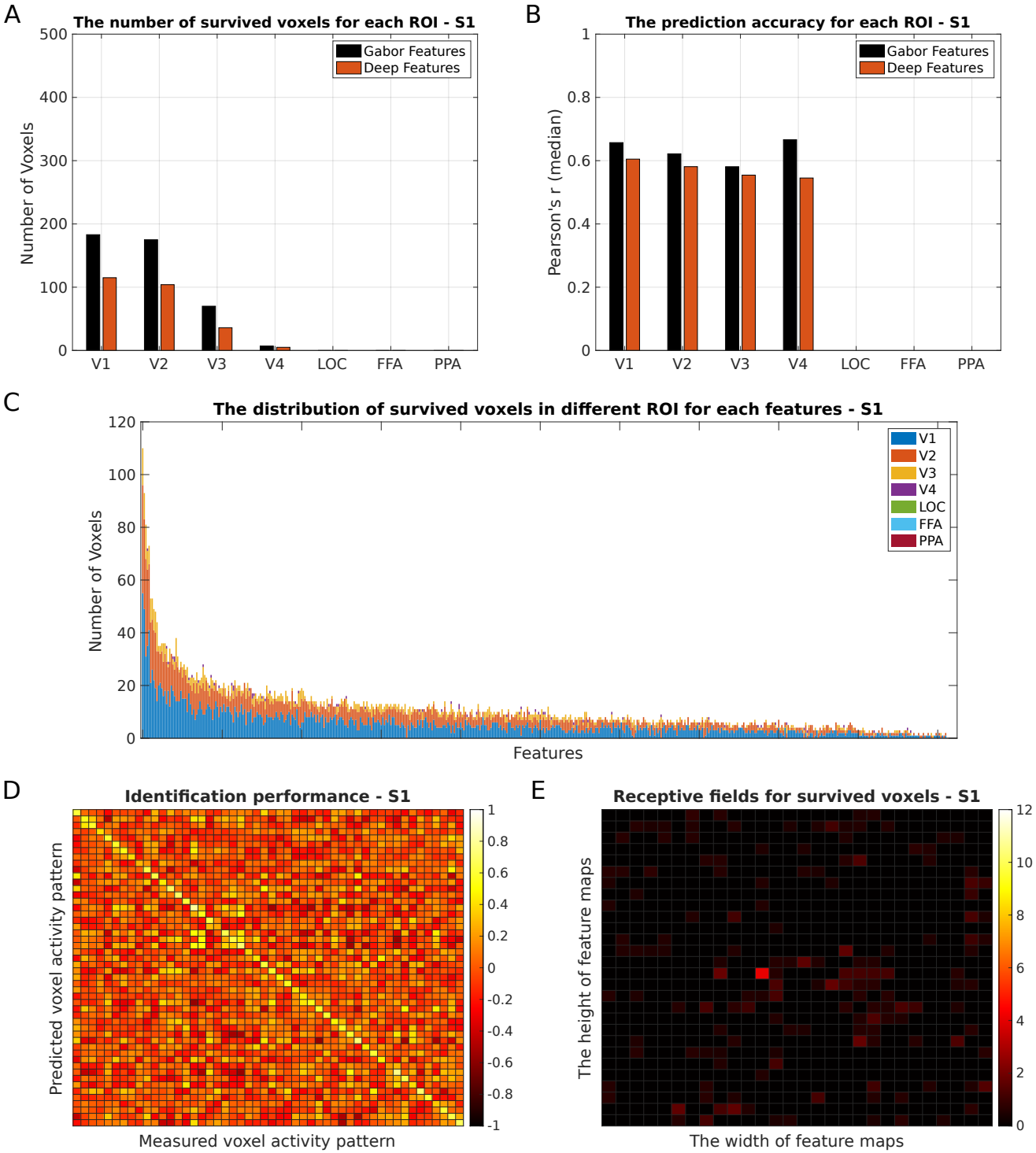

Fig. 2. The result of encoding models based on the deep features for S1. (A) The number of survived voxels for each ROI. (B) The prediction accuracy for each ROI (the median of Pearson's correlation coefficients of all survived voxels in each ROI). (C) The distribution of survived voxels in different ROI for each features. The features were ranked according to the number of voxels its related. (D) The decoding performance for S1 (identifying stimuli from measured brain activity using the testing set). (E) The distribution of receptive fields for survived voxels on the feature map. The value of each location equaled the sum of prediction accuracy (r) for all survived voxels located in that location.

V3, V4, LOC, FFA, and PPA were 0.69, 0.65, 0.61, 0.57, 0.55, 0.56, and 0.54 for S4; The prediction accuracies of the V1, V2, V3, V4, LOC, and PPA were 0.63, 0.60, 0.58, 0.54, 0.50, and 0.56 for S5.

Because of individual differences in brain structure and function, the pattern of prediction accuracy across ROIs was different among subjects. But we still observed some clear common trends: the features of the conv4_2 layer of the VGG19, which were selected as the semantic content representation of an image in the NST algorithm, mainly modulated the activity of voxels in the early visual areas (V1, V2, and V3). First, most of the survived voxels located in the early visual areas for each subject, and the number of survived voxels in other ROIs (V4, LOC, FFA, and PPA) are relatively few or just zero; Second, the prediction accuracy for early visual areas were slightly higher than other ROIs.

The survived models could be used to decode stimuli from the measured brain activity—image identification using the testing sets. The identification accuracies of 5 subjects (Fig. 2.D for S1) were 92% (46/50), 90% (45/50), 100%, 100%, and 92% (46/50). After checked all the identification errors, we founded that there were some common mistakes among different subjects. All the 4 images (No.17, No.19, No.41, and No.44) that were incorrectly identified by the encoding models of S5 were also incorrectly identified in S2, and three of them (No.19, No.41 and No.44) were incorrectly identified in S1 too. The encoding models of S1 made the same mistake as the models of S2, which identified the No.41 image as the No.42 image. And the encoding models of S2 made the same mistake as the models of S5, which identified the No.44 image as the No.26 image and the No.17 image as the No.22 image (For copyright reasons, we can not show the actual images).

Because Lasso regresssion enables feature selection, the survived models also described the relationship between features (X) and voxel responses (y) through regression coefficients. From the perspective of voxels, we calculated the number of features each ROI related (median) and found no common trend among subjects. From the perspective of features, we calculated the number of voxels each feature related and analyzed the location distribution of these voxels in different ROIs. After ranked features according to the number of voxels its related, we found that the deep features were mainly related to the voxels in the early visual areas (Fig. 2.C for S1). And we calculated Pearson's correlation coefficients of ranked feature index between each subject pair to examine if there were similar patterns among subjects. The result showed that there was no significant correlation. But if only considered the top-5 features, we found that most of the features were same among subjects (Table 1).

Table 1. Top-5 features (Index)

Feature Order	Subject1	Subject2	Subject3	Subject4	Subject5
1	383	250	250	383	383
2	60	60	383	60	250
3	289	383	60	250	60
4	250	289	355	448	482
5	355	198	289	355	289

The features (X) of each survived model corresponded to a spatial location on the feature map (we only used the features from one spatial location of the feature map as X), which could be seen as the center of population receptive field of the related voxel (y). So we could visualize the distribution of receptive fields of survived voxels for each subject (Fig. 2.E for S1). The result showed that survived voxels distributed widely on the feature map and there was a slight trend that some of voxels clustered near the center of the feature map.

The deep features may contained more information about stimuli than Gabor features. We also constructed encoding models based on Gabor features and compared it with encoding models using the deep features. Like the encoding models using the deep features, few voxels survived in the imagery sets when encoding models used gabor features (1 for S2 and 1 for S4). So we only compared two different types of encoding models using the test sets. From the perspective of individual voxel, there were more voxels survived in the early visual areas with encoding models based on Gabor features for all subjects(Fig. 2.A for S1). And the prediction accuracy of the early visual areas (the median of Pearson's correlation coefficients of all survived voxels in each ROI) was also higher for all subjects when encoding models used Gabor features (Fig. 2.B for S1). It implied that, relative to the deep features, Gabor features were preferentially represented by the early visual areas. From the perspective of activity pattern of voxels (decoding performance), however, the identification performances were better for all subjects when encoding models used the deep features (Fig 3).

The better identification performance of encoding models using the deep features could be due to the survived voxels in the higher visual areas, so we excluded survived voxels not in the early visual areas for both models and compared identification performance again. The results showed that, for all subjects, there were more survived voxels when encoding models used Gabor features (Fig. 4.B and Fig. 4.C for S1) but the identification performances were still better when encoding models used the deep features (Fig. 3). Then we chosed voxels simultaneously survived in the early visual areas for both models and found that the identification performances were better when encoding models used the deep features for all subjects (Fig. 3). Further analyses on these voxels, we observed a common trend among all subjects—the prediction accuracies of shared voxels showed a positive correlation between two types of models and most of voxels were better predicted when encoding models used Gabor features (Fig. 4.A for S1). Finally, we only chosed voxels which were better predicted by encoding models using Gabor features from shared voxels to compare the identification performance between two types of models. The result did not change—the identification performances were better when encoding models used the deep features for all subjects (Fig. 3), which may suggest that there was additional information in the deep features.

The representation of the semantic content of an image did related to the semantics of the image and also preserved visual details of the image to some extent. In accordance with the analysis using encoding models, the RSA was also based on individuals (Fig. 5 for S1). The RDMs from the pretrained VGG19 were significantly correlated with the RDM of the layer conv4_2. For all subjects, the RDM of the layer conv5_4 (the last convolutional layer) showed the strongest correlation ($r_s = 0.49$) and the RDM of the layer fc2 (the last

DRAFT

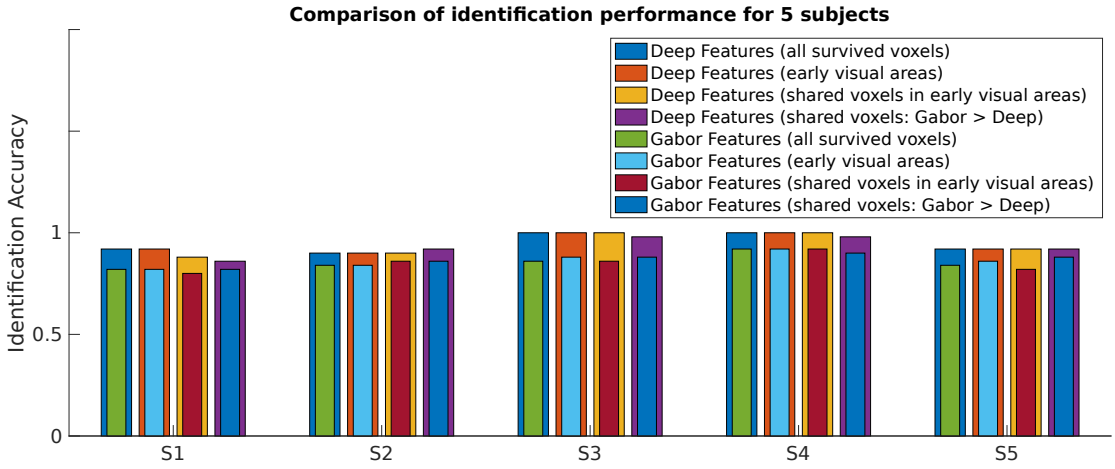

Fig. 3. The comparison of identification performance for 5 subjects.

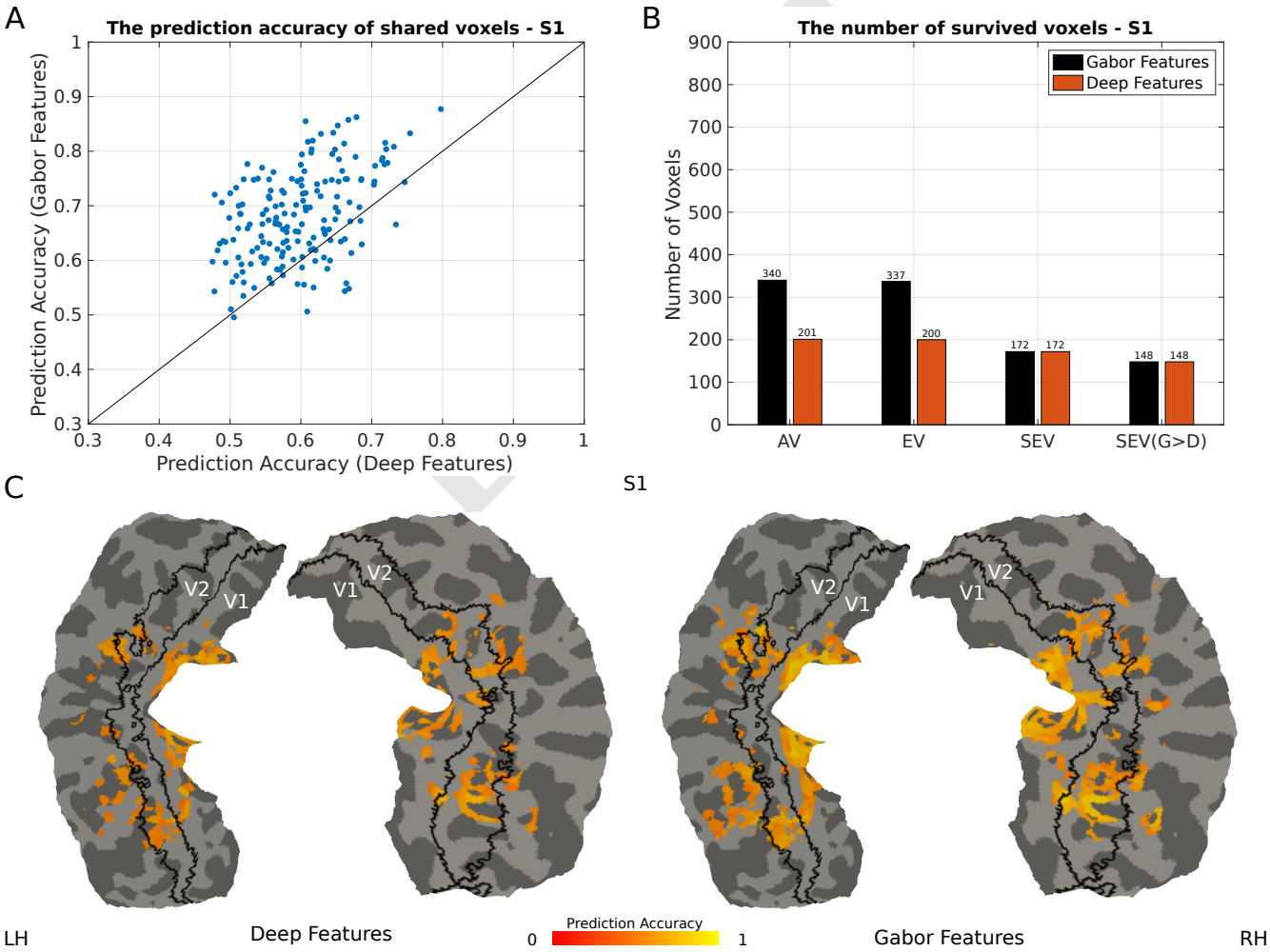

Fig. 4. The comparison of two types of encoding models for S1. (A) The prediction accuracy of shared voxels (voxels simultaneously survived in the early visual areas for both models) for S1. (B) The number of survived voxels for S1. AV = all survived voxels, EV = all survived voxels in the early visual areas, SEV = voxels simultaneously survived in the early visual areas for both models, SEV(G>D) = voxels which were better predicted by encoding models using Gabor features in SEV. (C) Prediction accuracy of survived voxels in the early visual areas mapped on the occipital surface. Surface reconstruction and flattening were performed using FreeSurfer (https://surfer.nmr.mgh.harvard.edu/).

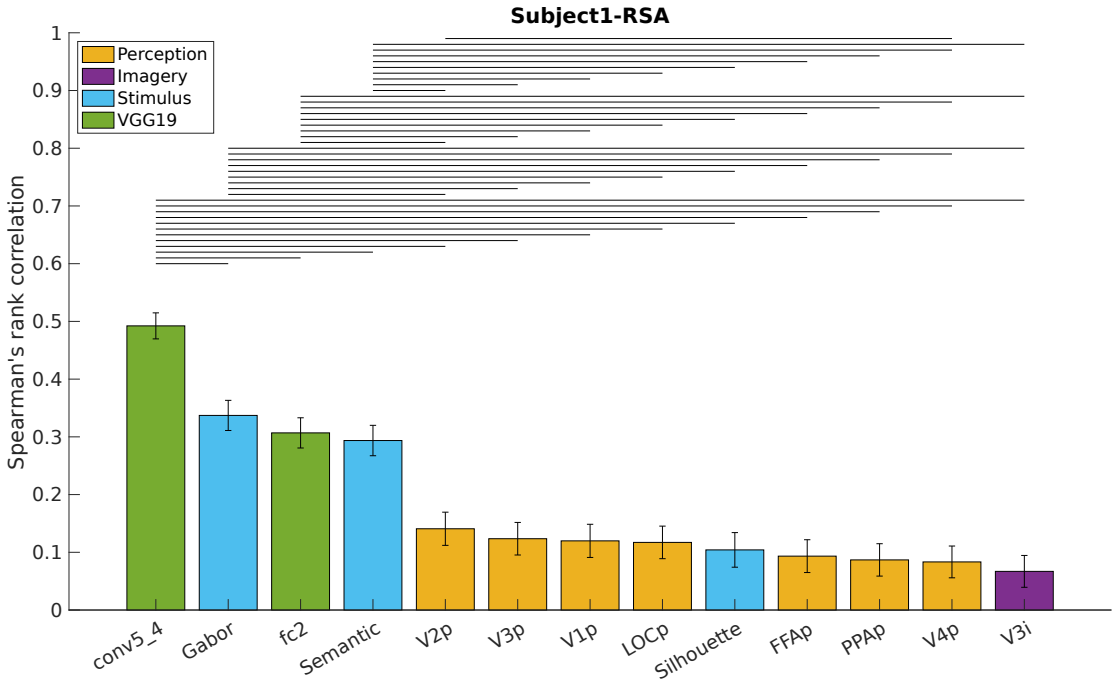

Fig. 5. The result of RSA for S1.

fully connected layer before SoftMax layer) was in the second echelon among all candidate RDMs ($r_s = 0.31$). This was a reasonable result given that all three RDMs derived from the same model (VGG19) and the layer conv5_4 was more similar to the layer conv4_2 than the layer fc2.

The RDMs from the stimuli were significantly correlated with the RDM of the layer conv4_2, too. For all subjects, the RDM of Gabor features ($r_s = 0.34$) and the RDM of semantics ($r_s = 0.29$) were both in the position of the second echelon. It implied that the representation of the layer conv4_2 did relate to the semantics of stimuli and also preserved visual details of stimuli to some extent. The RDM of silhouette was in the position of the lowest echelon for all subjects ($r_s = 0.10$). Because the silhouette of an object provided a limited description of the specific shape of the object, the representation of the layer conv4_2 also related to the specific shape of stimuli.

The situation of the RDMs from brain activity was more complex. For each subject, all 7 RDMs from the testing sets were significantly correlated with the RDM of the layer conv4_2. Although there were some individual differences about the relative position of these RDMs, the RDMs from the early visual areas which were in the second or third echelon roughly showed stronger correlation than those from the higher visual areas. In contrast, there were few RDMs from the imagery sets significantly related to the RDM of the layer conv4_2 (V3 for S1, PPA for S2, PPA for S3, V3, V4, LOC, and FFA for S5). And all of these RDMs were in the position of the lowest echelon. This result was in line with the result of encoding models to some extent.

Discussion

The impenetrability of DNNs reduced the explanatory power of studies which used DNNs as computational models to explore the biological brain. Inspired by NST, we circumvented this problem by using deep features that were given clear meaning—the representation of the semantic content of an image. Using encoding models, we quantitatively showed that the deep features which represented the semantic content of an image mainly modulated the activity of voxels in the early visual areas. It was a surprise that the semantics-related features mainly modulated the voxels in the early visual areas rather than those in the higher visual areas. Then we compared encoding models using the deep features with encoding models using Gabor features which have been proved to modulate the activity of voxels in the early visual areas (41) and found that the deep features may contained more information about stimuli than Gabor features. These results naturally led us to another question: what the representation of the semantic content of an image really is? The result of RSA showed that, the representation of the semantic content of an image did related to the semantics of the image and also preserved visual details of the image to some extent. It implied that the representation of the semantic content of an image might be a hybrid form—both in propositional format and depictive format.

How could the format of a representation be both propositional and depictive? The question of how the information is represented in our brain had been discussed for many years, which was known as the imagery debate (49). At the heart of the debate was whether all information is represented in a symbolic, propositional format. Convergent evidence from empirical studies of mental imagery suggested that information can be represented in a pictorial, depictive format (50). The existence of the depictve format of information ended the imagery debate but also raised new questions: how many formats can the brain use and what is the relationship between these formats and the propositional format? With the development

DRAFT

of theories of grounded cognition, the dominant position of the propositional format in cognition is being challenged. From the perspective of grounded cognition, there were no amodal symbols in our brain that were independent of the modal representation and all cognitive phenomena were ultimately grounded in modal simulations, bodily states, and situated action, which was supported by many researches on perception, memory, language, thought, social cognition, and development (45–47). This view emphasized the key role of the depictive, modality-specific representation in cognition and denied the independent existence of the symbolic, propositional representation, which was clearly articulated by Comenius from several hundred years ago—"things are essential, words only accidental; things are the body, words but the garment; things are the kernel, words the shell and husk. Both should be presented to the intellect at the same time, but particularly the things, since they are as much objects of understanding as is language" (51).

From this view, the representation of information in our brain is essentially depictive and can implement symbolic functions naturally. This is in line with our result to some extent. On the one hand, the representation of the semantic content of an image (the feature maps of the layer conv4_2) was essentially depictive. This was because the feature maps extracted from the convolutional layer of the VGG19 naturally preserved the topology of the original image. Besides, The reuslt of RSA also showed that it preserved visual details of the image. On the other hand, this representation did reflect the semantics of the image in some degree. To the best of our knowledge, it is the first time that the existence of such representation in our brain is quantitatively proved, at least in the early visual areas.

Unlike the previous study (14), we did not observe that encoding models which were trained using the perceptual data could successfully predict voxel responses from the imagery data. This could be due to different experimental tasks. In the study of (14), subjects were asked to imagine particular artworks, such as "Betty" by Gerhard Richter and "Horse Bath" by Odd Nerrdum. In the study of (16), which provided data for this paper, subjects were asked to imagine as many object images as possible from concrete nouns, such as leopard and swan. The difference between two tasks was whether the imagery had a particular content. For example, when you were asked to draw your cat or dog, what you drew must be a particular cat or dog; but when you were asked to draw a cat or dog, you could draw any cat or dog, even Hello Kitty or Snoopy. Because of the individualization and arbitrariness of the imagery in our study, it seems reasonable that our result was not consistent with the previous study and could not address the issue of the relationship between perception and imagery.

In addition to the obvious individual divergences in encoding mechanisms, our result showed that there were some common mechanisms among subjects (common mistakes in image identification and similarity of the top 5 features). In contrast to the symbolic, propositional representations, the depictive, modality-specific representations of information were grounded in the modalities, the body, and the environment. So they were highly personal and changed from time to time. This was a key difference between the grounded cognition and traditional cognitive theories and could be used to explain individual divergences in cognition. Meanwhile, we did share a common physical and environmental basis, which was also reflected in cognitive process and made communication possible. This may explain the existence of the common mechanisms.

In our study, we quantitatively showed that the deep features which represented the semantic content of an image mainly modulated the activity of voxels in the early visual areas and these features were essentially depictive but also propositional. This result implied that some depictive representation of an object in our brain can naturally reflect semantics of the object to some extent and this phenomena can be found in the early visual areas, which provided empirical evidence to the core viewpoint of the ground cognition. In fact, there was another theory also addressing the relationship between the propositional representation and the depictive representation of information in our brain—the dual coding theory (52), which emphasized the beneficial effects of the depictive representation of information on cognition (concreteness) and suggested that the two types of representations are independent from each other in our brain (Paivio believed that there were two distinct subsystems in our brain specialized for dealing with different types of representations). Both theories admitted the association between the two types of representations but disagreed with each other about the relationship between the two types of representations. From our result, we tend to support the monistic view of the two types of representations. But we also noticed that our result only involved the early visual areas which were in the early stages of the visual ventral stream. So How does the depictive representation of an object change along the visual ventral stream to make object recognition possible—whether the independent propositional representation will eventually appear—is not clear. This question needs further studies.

In addition, our study also provided complementary evidence to NST. Although NST is an active area of computer vision and lead to many successful industrial applications, it still can not clearly define what is the semantic content of an image and what is the style of an image. According to our result, the feature map of the layer conv4_2 of VGG19, which was selected as the representation of the semantic content of an image from the original NST algorithm, did related to the semantics of the image and also preserved visual details of the image to some extent. It implied that the semantic content of an image may represent the semantics of the image in a concrete way.

Materials and Methods

Data. We used the fmri data that was originally published in (16), which can be downloaded from https://github.com/KamitaniLab/GenericObjectDecoding. The data was obtained from two fMRI experiments for each of 5 subjects: an image presentation experiment and an imagery experiment. There were two sessions in the image presentation experiment—the training session and the testing session. In the training session, subjects viewed 1200 images from 150 categories (8 images from each category) as each image presented once. In the testing session, subjects viewed 50 images from 50 categories (one image from each category) as each image presented 35 times. In the imagery experiment, subjects were asked to imagine about 50 nouns from 50 categories (one noun from each category) as each noun presented 10 times. The categories used in the imagery experiment were the same as those in the testing session, which were not used in the training session. All the images (1250 natural

DRAFT

images with the resolution of $512 \times 512 \times 3$) and the corresponding categories (200) were collected from ImageNet (53). In addition, there were standard retinotopic mapping experiment and functional localizer experiment to identify lower visual areas (V1, V2, V3, and V4) and higher visual areas (LOC, FFA and PPA) for each participant. The details of the experimental design, MRI acquisition protocol, and preprocessing of the fMRI data could be found in (16).

Before further analysis, we averaged the repeated trials in the testing session and the imagery experiment. First, we standardized the fMRI data from the training session. The mean and standard deviation of the training set were then used to standardize the testing sets (from the testing session) and the imagery sets (from the imagery experiment). After that, we performed trial-averaging for the testing sets and the imagery sets to improve the signal-noise ratio. Because of trial-averaging, there were statistical difference between the training set and the other two. So we rescaled the averaged data by a factor of $\sqrt{n}$ where n is the number of trials averaged (17).

Encoding Models Based On Deep Features. To probe how does the brain represent the semantic content of an image, we used voxel-wise encoding models. Such models were constructed separately for each voxel in each individual, so our analysis was also individual-based. There were two steps to construct voxel-wise encoding models (39): the first step was a nonlinear transformation from a stimulus space to a feature space; the second step was a linear transformation from the feature space to a voxel space.

As the first step, we got deep features which represented the semantic content of an image. We employed the pre-trained VGG19 model based on the open source machine learning framework of PyTorch (54) to extract the feature maps of the conv4_2 layer as the image content representation. After image preprocessing (such as image scaling and cropping, more details can be found from https://pytorch.org/hub/pytorch_vision_vgg/), images from the training data session and the testing data session were fed into the VGG19 model and the feature maps of the conv4_2 layer were extracted, respectively. As a result, the size of the training feature maps was [1200, 512, 28, 28] ([the number of images, the number of kernels, the height of feature map, the width of feature map]), and the size of the testing feature maps was [50, 512, 28, 28].

As the second step, we constructed linear regression models to predict the brain activity evoked by an image from the features which represented the semantic content of the same image. For each voxel, the model can be expressed by

$$y = X\beta + \epsilon \qquad [1]$$

Where y is a measured voxel response and X is features which elicited the response of the voxel.

Just as each neuron has its own receptive field, each voxel has its own population receptive field (55). It means that a voxel only responds to the features in its population receptive field. So there was no need to put all the features into the model (The number of features for each image is 401408, It will pose a problem known as the curse of dimensionality if we put all the features into the model). And the features we extracted from the VGG19 model were naturally organized into feature maps that preserved the topology of stimuli. For example, the size of the feature maps for each image was $512 \times 28 \times 28$. It could be seen as 784 (28×28) spatial locations with 512 features at each spatial location. The spatial arrangement of 784 locations (28×28) preserved the topology of the original image (512×512). According to the model of the population receptive field, receptive fields are center-surround organized and features at the center of a receptive field make the largest contribution to the activity measured in the voxel (56). So we only used 512 features at one of the 784 locations as X for each voxel (To simplify the model and reduce computation time, we ignored the surround of receptive fields and assumed that each location is a candidate for the center of a receptive field). To find the best center for each voxel, we constructed separate linear regression models for each location/voxel combination.

We used the training data to fit models. For each model, X was the features from one of the 784 locations represented by a 1200×512 matrix (1200 images), y was a measured voxel response represented by a 1200×1 matrix. It was reasonable to assume that a voxel only responded to a fraction of the X. So we estimated regression coefficients of each model using lasso regression:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \qquad [2]$$

where lambda is a complexity parameter that controls the amount of regularization. To accelerate model fitting, we used the function "lasso_gpu" from a MATLAB package developed by (57), which can be efficiently implemented in parallel on a GPU. The optimal value of lambda was selected from a lambda sequence with 10 lambdas: $2^{-2}, 2^{-3}, 2^{-4}, \cdots, 2^{-11}$ (the first lambda was set using the function "calculate_lambda_start" from the same MATLAB package and the lambda range was set to guarantee that the last lambda is not chosen as the best one). We chose the best lambda and the best location for each voxel using five-fold cross-validation with the coefficient of determination (R^2).

Model Evaluation. Once fitted, encoding models were evaluated using the testing sets and the imagery sets, respectively. For each voxel, we defined the model's prediction accuracy as the Pearson's correlation coefficient (r) between the measured voxel response and the response predicted by the model. The significance of the correlation was assessed by a permutation test with 10000 permutations (Bonferroni correction for the number of voxels, and $p < 0.05$). For each ROI, we used the number of survived voxels within the ROI and the meidan r as measurements for prediction accuracy.

Another measurement for prediction accuracy was decoding performance—identifying stimuli from measured brain activity (41). First, we used survived models to predict the voxel activity pattern from the test feature maps for each of the 50 stimuli. Second, we calculated the Pearson's correlation coefficient between the predicted voxel activity pattern and the measured voxel activity pattern for each stimulus/stimulus combination. The stimulus whose predicted voxel activity pattern was most correlated with the measured voxel activity pattern of itself was regarded as correct decoding. We defined the identification accuracy as the percentage of stimuli that are correctly identified from the testing data or the imagery data.

Model Analysis. After model evaluation, we explored the relationship between features and voxels through regression coefficients of survived models. Because of the lasso regression, some of regression coefficients in each model were set to zero automatically. The nonzero coefficients indicated that which features were related to the activity of a voxel and how are they related. For each subject, we described the relationship between features and voxels from two complementary perspective—the perspective of features and the perspective of voxels. From the perspective of features, we used the number of voxels each feature related and the location distribution of these voxels in different ROIs to analyse the property of different features. From the perspective of voxels, we used the number of features each voxel related and the location of the voxel to analyse the property of different ROIs.

Furthermore, the survived models also provided information about the population receptive fields of voxels. For each model, the X was selected from one of the 784 locations (28×28). The location of the X could be seen as the center of the population receptive field of the corresponding voxel on the feature maps. We used heatmap to visualize and explore the distribution of the receptive fields of survived voxels for each subject.

Encoding Models Based On Gabor Features. To compare with encoding models using the deep features, we also constructed encoding models based on Gabor features. We got gabor features of stimuli according to the method of (58): Firstly, a Gabor Wavelet Pyramid (GWP, 59) model was used to get original Gabor features from stimuli (six spatial frequencies: 1, 2, 4, 8, 16, and 32 cycles/FOV; eight orientations: 0°, 22.5°, 45°, . . . , and 157.5°; and two phases: 0° and 90°; The FOV covered full of a image and all images were downsampled to 128×128 pixels); Secondly, the absolute values of the projections of each quadrature wavelet pair were averaged to get the contrast energies; Thirdly, the contrast energies were normalized to linearize the relationship between contrast energies and voxel responses (each contrast energy was divided by the sum of the contrast energy and the median of all contrast energies in the training set which were at the same position and the same

DRAFT

orientation); Fourthly, the normalized contrast energies at the same position (eight orientations) were averaged to reduce the dimension of features; Lastly, the average luminance of stimuli were also added into Gabor features. As a result, each stimulus had 1366 features. After that we used the same method as described above to construct linear regression models from Gabor features to voxel responses.

Representational Similarity Analysis. To probe the representation property of the layer conv4_2, we used RSA, which characterized the representational geometry of a set of stimuli in a brain region or computational model by representational distance matrix (RDM) and compared RDMs to explore the representational similarity between different brain regions or brain regions and computation models. Three types of RDMs (the candidate RDMs) were constructed to compare with the RDM of the layer conv4_2 (the reference RDM).

The first type were RDMs derived from the pre-trained VGG19. There were two RDMs—the RDM of the layer conv5_4 and the RDM of the layer fc2. As the RDM of the layer conv4_2, the two candidate RDMs were constructed using the corresponding feature maps extracted from the VGG19 by the 50 images from the testing set. We selected the correlation distance (1 minus the linear correlation between each pair of feature maps) as the measurement of representational dissimilarity to construct each RDM.

The second type were RDMs derived from brain activity. We used the measured brain activity from the testing sets and the imagery sets to constructed RDMs, respectively. Because the categories of the stimuli (nouns) in the imagery data were the same as the categories of the stimuli (images) in the test data, the RDMs from the imagery data could be treated as RDMs using the same set of stimuli—the 50 images from the test data. There were 14 RDMs for each subject, 7 RDMs from the test data for each ROI and 7 RDMs from the imagery data for each ROI (V1, V2, V3, V4, LOC, FFA, and PPA).

The third type were RDMs derived from stimuli (50 images from the testing set) directly. There were three RDMs—the RDM of Gabor features, the RDM of silhoutee, and the RDM of semantics. The RDM of Gabor features was constructed using the Gabor features of images extracted from the GWP model. To construct the RDM of silhoutte, we converted images to silhouettes (binary images in which each figure pixel is 0 and each background pixel is 1) and calculated the correlation distance between each pair of silhouettes. To construct the RDM of semantics, we calculated the semantic distance between each pair of images. We used the function "path_similarity" from the Natural Language Toolkit library (60) to calculate how similar two categories of images are (semantic similarity), which based on the WordNet (61). The score returned from the function was in range 0 to 1, so we converted the semantic similarity to the semantic distance by subtracting the score from 1.

In total 19 candidate RDMs were constructed to compare with the reference RDM for each subject and the Spearman's rank correlation coefficient was selected to measure the similarity between each candidate RDM and the reference RDM. After that, we performed statistical inference to answer two questions—whether a candidate RDM and the reference RDM were significantly correlated (by permutation test with 10000 permutations) and whether the correlation between a candidate RDM and the reference RDM was significantly different from the correlation between another candidate RDM and the reference RDM (by bootstrap test with 1000 replications). For each test, FDR was applied for multiple comparison correction (62). All calculations were done using MATLAB 2020a and Python 3.7 on a Linux (Ubuntu 18.04 LTS) desktop with a Geforce GTX 1660 Ti graphics card (6 Gb of VRAM).

ACKNOWLEDGMENTS. This work was supported by the National Natural Science Foundation of China (No. 31600907).

1. RM Cichy, D Kaiser, Deep neural networks as scientific models. *Trends Cogn. Sci.* **23**, 305–317 (2019).
2. JI Glaser, AS Benjamin, R Farhoodi, KP Kording, The roles of supervised machine learning in systems neuroscience. *Prog. Neurobiol.* **175**, 126–137 (2019).
3. N Kriegeskorte, Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
4. GW Lindsay, Convolutional neural networks as a model of the visual system: Past, present, and future. *J. Cogn. Neurosci.*, 1–15 (2020).
5. BA Richards, et al., A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
6. JJ DiCarlo, DD Cox, Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
7. JJ DiCarlo, D Zoccolan, NC Rust, How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
8. SA Cadena, et al., Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Comput. Biol.* **15**, e1006897 (2019).
9. SM Khaligh-Razavi, N Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
10. DLK Yamins, et al., Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* **111**, 8619–8624 (2014).
11. DLK Yamins, JJ DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
12. M Eickenberg, A Gramfort, G Varoquaux, B Thirion, Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* **152**, 184–194 (2017).
13. U Guclu, MAJ van Gerven, Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
14. T Naselaris, CA Olman, DE Stansbury, K Ugurbil, JL Gallant, A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage* **105**, 215–228 (2015).
15. K Seeliger, et al., Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage* **180**, 253–266 (2018).
16. T Horikawa, Y Kamitani, Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* **8** (2017).
17. G Shen, K Dwivedi, K Majima, T Horikawa, Y Kamitani, End-to-end deep image reconstruction from human brain activity. *Front. Comput. Neurosci.* **13** (2019).
18. H Wen, et al., Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* **28**, 4136–4160 (2017).
19. DGT Barrett, AS Morcos, JH Macke, Analyzing biological and artificial neural networks: challenges with opportunities for synergy? (2018).
20. D Hassabis, D Kumaran, C Summerfield, M Botvinick, Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).
21. KN Kay, Principles for models of neural information processing. *NeuroImage* **180**, 101–109 (2018).
22. C Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
23. A Krizhevsky, I Sutskever, GE Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
24. K Simonyan, A Zisserman, Very deep convolutional networks for large-scale image recognition. (2014).
25. B Zhou, D Bau, A Oliva, A Torralba, Interpreting deep visual representations via network dissection. *IEEE Transactions on Pattern Analysis Mach. Intell.* **41**, 2131–2145 (2019).
26. A Mahendran, A Vedaldi, Understanding deep image representations by inverting them in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (IEEE), (2015).
27. A Nguyen, J Yosinski, J Clune, Understanding neural networks via feature visualization: A survey in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. (Springer International Publishing), pp. 55–76 (2019).
28. C Olah, A Mordvintsev, L Schubert, Feature visualization. *Distill* **2** (2017).
29. J Yosinski, J Clune, A Nguyen, T Fuchs, H Lipson, Understanding neural networks through deep visualization. (2015).
30. MD Zeiler, R Fergus, Visualizing and understanding convolutional networks in *Computer Vision – ECCV 2014*. (Springer International Publishing), pp. 818–833 (2014).
31. K Kar, J Kubilius, K Schmidt, EB Issa, JJ DiCarlo, Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
32. K Han, et al., Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage* **198**, 125–136 (2019).
33. C Zhuang, et al., Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci.* **118**, e2014196118 (2021).
34. R Geirhos, et al., Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. (2018).
35. Y Jing, et al., Neural style transfer: A review. *IEEE Transactions on Vis. Comput. Graph.* **26**, 3365–3385 (2020).
36. LA Gatys, AS Ecker, M Bethge, Image style transfer using convolutional neural networks in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (IEEE), (2016).
37. LA Gatys, AS Ecker, M Bethge, Texture and art with deep neural networks. *Curr. Opin. Neurobiol.* **46**, 178–186 (2017).
38. N Kriegeskorte, PK Douglas, Interpreting encoding and decoding models. *Curr. Opin. Neurobiol.* **55**, 167–179 (2019).
39. T Naselaris, KN Kay, S Nishimoto, JL Gallant, Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
40. MA van Gerven, A primer on encoding models in sensory neuroscience. *J. Math. Psychol.* **76**, 172–183 (2017).
41. KN Kay, T Naselaris, RJ Prenger, JL Gallant, Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
42. N Kriegeskorte, Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* (2008).
43. N Kriegeskorte, RA Kievit, Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).
44. H Nili, et al., A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).

DRAFT

45. LW Barsalou, Grounded cognition. *Annu. Rev. Psychol.* **59**, 617–645 (2008).
46. LW Barsalou, Grounded cognition: Past, present, and future. *Top. Cogn. Sci.* **2**, 716–724 (2010).
47. LW Barsalou, Challenges and opportunities for grounding cognition. *J. Cogn.* **3** (2020).
48. DH Hubel, TN Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The J. Physiol.* **160**, 106–154 (1962).
49. J Pearson, SM Kosslyn, The heterogeneity of mental representation: Ending the imagery debate. *Proc. Natl. Acad. Sci.* **112**, 10089–10092 (2015).
50. J Pearson, The human imagination: the cognitive neuroscience of visual mental imagery. *Nat. Rev. Neurosci.* **20**, 624–634 (2019).
51. A Paivio, *Mind and its evolution : a dual coding theoretical approach.* (L. Erlbaum Associates, Mahwah, N.J), (2007).
52. JM Clark, A Paivio, Dual coding theory and education. *Educ. Psychol. Rev.* **3**, 149–210 (1991).
53. J Deng, et al., ImageNet: A large-scale hierarchical image database in *2009 IEEE Conference on Computer Vision and Pattern Recognition.* (IEEE), (2009).
54. A Paszke, et al., Pytorch: An imperative style, high-performance deep learning library. (2019).
55. SO Dumoulin, BA Wandell, Population receptive field estimates in human visual cortex. *NeuroImage* **39**, 647–660 (2008).
56. G St-Yves, T Naselaris, The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage* **180**, 188–202 (2018).
57. H Mohr, H Ruge, Fast estimation of l1-regularized linear models in the mass-univariate setting. *Neuroinformatics* (2020).
58. C Li, J Xu, B Liu, Decoding natural images from evoked brain activities using encoding models with invertible mapping. *Neural Networks* **105**, 227–235 (2018).
59. G Gaziv, Gabor wavelet pyramid (https://www.mathworks.com/matlabcentral/fileexchange/60088-gabor-wavelet-pyramid) (MATLAB Central File Exchange) (2021).
60. S Bird, E Klein, E Loper, *Natural Language Processing with Python.* (O'Reilly UK Ltd.), (2009).
61. GA Miller, Wordnet: A lexical database for english. *Commun. ACM* **38**, 39–41 (1995).
62. Y Benjamini, Y Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B (Methodological)* **57**, 289–300 (1995).

DRAFT