Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Neural encoding with unsupervised spiking convolutional neural network

**Chong Wang**
**Hongmei Yan**
**Wei Huang**
**Wei Sheng**
**Yuting Wang**
**Yun-Shuang Fan**
**Tao Liu**
**Ting Zou**
**Rong Li**
**Huafu Chen** ( chenhf@uestc.edu.cn )
University of Electronic Science and Technology of China https://orcid.org/0000-0002-4062-4753

**Article**

**Keywords:**

**Posted Date:** February 20th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2555652/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Version of Record:** A version of this preprint was published at Communications Biology on August 28th, 2023. See the published version at https://doi.org/10.1038/s42003-023-05257-4.

# Abstract

Accurately predicting the brain responses to various stimuli is big challenge in neuroscience. Although recent fMRI studies have made significant breakthroughs in neural encoding using convolutional neural networks (CNNs), there remains critical gaps between the computational rules of traditional artificial neurons and real biological neurons. Here, we present a spiking CNN (SCNN)-based framework to accomplish neural encoding in a more biologically plausible way. The framework extracts visual features of the image stimuli using the unsupervised SCNN and predicts fMRI responses from the SCNN features using a receptive field-based regression algorithm. Experimental results on both handwritten characters and natural images demonstrate that our approach can achieve surprisingly good encoding performance and be used for realizing “brain reading” tasks such as image reconstruction and image identification. Our work suggests that SNN can serve as a promising tool for neural encoding.


# Introduction

The goal of neural encoding is to predict how the brain responds to the outside inputs. It provides an effective approach to explore the brain mechanism for processing sensory information and can form the basis of brain–computer interface systems. Visual perception, as one of the most important ways for us to receive outside information, has been the major target of neural encoding. Over the past two decades, scientists have made remarkable progress in vision-based neural encoding[1–4] with the development of non-invasive brain imaging techniques, such as functional magnetic resonance imaging (fMRI), thereby making it a hot topic of neuroscience.

The common steps for vision-based encoding include feature extraction and response prediction[5]. Feature extraction aims to produce visual features of the stimuli by simulating the visual cortex. A feature extractor that approximates the real visual mechanisms is essential for encoding. Response prediction aims to make voxel-wise predictions of the fMRI responses based on the extracted visual features. This step is usually accomplished by linear regression[6] because the relationship between the features and the responses should be as simple as possible. Previous studies demonstrated that the early visual cortex processes information in a way similar to Gabor wavelets[7–9]. Inspired by this finding, Gabor filter-based encoding models have been proposed and successfully applied in image identification and movie reconstruction tasks[1,3]. In recent years, convolutional neural networks (CNNs) have drawn extensive attention because of their remarkable achievements in the field of computer vision. Representational similarity analysis[10] results show that human visual cortex share similar hierarchical representations to CNNs[11,12]. Therefore, CNN-based encoding models began to be widely used and have achieved excellent performance[2,4,13,14]. Despite the great success of CNNs in the application of encoding, the differences between CNNs and the brain in processing visual information cannot be ignored [15].

In terms of computational mechanisms, the key distinction between the artificial neurons in CNNs and the biological neurons is that the former propagates continuous digital values, whereas the latter propagates

action potentials (spikes). The proposal of spiking neural network (SNNs), which was regarded as the third generation of neural networks[16], narrowed this difference. Unlike traditional artificial neural networks (ANNs), SNNs convey information by spike timing. In SNNs, each neuron integrates spikes from the previous layer and emits spikes to the next layer when its internal voltage reaches the threshold. The most commonly used learning algorithm of SNNs is the spike-timing-dependent plasticity (STDP)[17,18], which is an unsupervised method for weight update and has been uncovered in mammalian visual cortex[19–21]. Recent studies have applied STDP-based SNNs on object recognition and achieved considerable performance[22–24]. The advantage of SNNs in biological plausibility leads to their potential in neural encoding.

In this paper, we proposed a spiking CNN (SCNN)-based encoding framework to fill in the gap between CNNs and the realistic visual system. The encoding procedure consisted of three steps. First, we trained a SCNN using STDP algorithm to extract the visual features of the images. Second, we annotated the coordinates of each voxel's receptive field in the SNN feature maps based on the retinal topological properties of the visual cortex. That is, each voxel receives visual input from only one fixed location of the feature map. Third, we built linear regression models for each voxel to predict their responses from corresponding SNN features. The framework was tested using two publicly available image-fMRI datasets, namely, handwritten character[25] and natural image[1] datasets. Moreover, we realized two downstream decoding tasks (image reconstruction and image identification) based on our encoding models. We compared the encoding and decoding performance of our method with those of previous methods.

# Results

**Encoding performance on handwritten character dataset.** For this dataset, we first built the SCNN using the images in the TICH dataset (images in the test set were excluded, and 14854 images for the 6 characters were retained), which aimed to maximize the representation ability of the SCNN. Then, we trained the voxel-wise linear regression models with the fMRI data in the train set for each participant. The encoding performance was defined as the PCC between the predicted and measured responses to the test set images. Moreover, we compared our model with the CNN-based encoding model, in which the network architecture of CNN was constrained to be consistent with that of the SCNN (Supplementary Table 1). The CNN was trained with Adam optimizer[26] with a learning rate of 0.0001 for 50 epochs using TICH dataset, and the classification accuracy achieved 99% on the test set images. The subsequent encoding procedures for CNN were the same as SCNN. For each subject, 500 voxels with highest $R^2$ in the 3-fold cross-validation of the train set data were selected for the comparison. Figure 2a shows the prediction accuracies for SCNN and CNN-based encoding models, we found that the accuracies of SCNN on all of the three subjects were significantly higher than those of CNN ($p<10^{-13}$, one-tailed two sample t-test). This result suggests that the SCNN has greater potential than CNN in terms of encoding tasks.

The predictability of the activity of a voxel depends on its degree of involvement in the task. That is, if a voxel receives a large amount of stimulus information, then its fMRI activities will be easy to predict, and vice versa. To verify this, we visualized the distributions of the stimulus intensities and voxel receptive fields. As we annotated the receptive field for each voxel through the 3-fold cross-validation on the train set data, the top 100 voxels with the highest $R^2$ of each participant were adopted for the analysis. Figure 2b shows the mean stimulus intensities of the train set and the receptive fields of the selected voxels. We found that their spatial distribution patterns, which approximately obeyed Gaussian distributions along the X- axis and uniform distributions along the Y- axis, were quite similar. In other words, the receptive fields of these informative voxels tended to be distributed in locations with higher stimulus intensity. This finding also proves the effectiveness of our receptive field-based feature selection algorithm.

**Encoding performance on natural image dataset.** Compared with handwritten character images, natural images are more complex and closer to our daily vision. To validate whether our approach is feasible for encoding natural image stimuli, we trained and tested the proposed encoding model based on this dataset. Task-optimized CNN-based models are not applicable in this encoding task because the visual stimuli in this dataset have no category labels. We compared our approach with the Gabor wavelet pyramid (GWP) model proposed by Kay, et al. [1] and the brain-optimized CNN (GNet)[13,27]. Instead of classifying the input images, the CNN in GNet was trained for predicting the fMRI responses in end-to-end fashion. The architecture of GNet can be found in Supplementary Table 2, we trained the GNet independently for each visual area in each subject (a total of 6 models were trained). As shown in Table 1, the number of predictive voxels ($p < 0.05$, Bonferroni corrected) of the proposed model were larger than that of GWP, and next to the GNet. We noted that there were fewer predictable voxels in V3, which might be related to its lower signal-to-noise ratio[1]. Figure 3 shows the prediction accuracies for different models of the predictive voxels. For each visual areas, the accuracies of SCNN were significantly higher than those of GWP ($p<10^{-6}$, one-tailed two sample t-test) and have no significantly difference with GNet ($p>0.37$, two-tailed two sample t-test). These results indicate that the unsupervised SCNN-based encoding model is superior to GWP, and can even achieve comparable performance with the neural network that optimized with brain response as target.

Table 1
Comparison of the predictable voxels and the identification accuracies of GWP and SCNN based encoding models on natural image dataset.

| Subject | Visual area | Total number of voxels | Predictable voxels | | | Identification accuracy | | |
|---|---|---|---|---|---|---|---|---|
| | | | GWP | GNet | SCNN | GWP | GNet | **SCNN** |
| S1 | V1 | 1331 | 256 | 380 | 342 | 92% | | **96.67%** |
| | V2 | 2208 | 217 | 444 | 371 | | 90% | |
| | V3 | 1973 | 75 | 174 | 124 | | | |
| S2 | V1 | 1513 | 108 | 206 | 166 | 72% | | **90.83%** |
| | V2 | 1982 | 89 | 192 | 138 | | 73.33% | |
| | V3 | 1780 | 12 | 16 | 30 | | | |

**Image reconstruction result**. Based on the encoding model, we accomplished the image reconstruction task using the handwritten character dataset. We adopted the images of the six characters in the TICH dataset (excluding the test set images) as the prior image set and used the encoding model to produce the predicted fMRI responses for each image. Notably, only 500 voxels selected from the train set data were used for this task. For each image in the test set, we averaged the top 15 images of the prior image set whose predicted responses were most correlated to the observed responses as the reconstructed image. Figure 4a, b show some reconstruction examples. We obtained promising reconstruction results on every character, and the images that belong to the same character with different writing styles can be reconstructed. This result indicates that the reconstructions contain both category and structure information of the original stimuli. Moreover, we compared our reconstruction results with those of other methods, including CNN, SMLR[28], DGMM[29], Denoiser GAN[30], and D-VAE/GAN[31]. The reconstruction performance was evaluated using PCC and Structural Similarity Index (SSIM)[32]. As shown in Table 2, our approach achieved competitive or superior performance to the state-of-the-art methods.

Table 2
Reconstruction performance with different methods on handwritten character dataset.

| Methods | PCC | SSIM |
|---|---|---|
| CNN | 0.264±0.136 | 0.167±0.050 |
| SMLR[28] | 0.481±0.096 | 0.191±0.043 |
| DGMM[29] | 0.502±0.193 | 0.360±0.050 |
| Denoiser GAN[30] | 0.319±0.032 | 0.465±0.031 |
| D-VAE/GAN[31] | 0.740±0.020 | **0.587±0.019** |
| SCNN (Ours) | **0.759±0.100** | 0.548±0.174 |

The image reconstruction task has two hyper-parameters: the number of selected voxels and the number of averaged images. To estimate the effect of the hyper-parameters to the reconstruction, we investigated the reconstruction performance using the voxels with the number of 100, 200, 500, and 1000 and the images with the number of 1, 5, 10, 15, 20, 25, and 30. As shown in Fig. 4c, our approach achieved the highest performance on PCC with the voxel number of 500 and the image number of 15. The reconstruction performance on SSIM decreased with the increase in the image number and reached the highest with the voxel number of 500. A larger number of voxels contained more stimulus information but they also introduced a lot of noise. Likewise, a larger number of images made the reconstruction more realistic but also blurred the reconstructions.

**Image identification result.** The image identification task was accomplished using the natural image dataset. The encoding model was used to produce predicted fMRI responses for all the images in the test set. The images perceived by the participants were identified by matching the measured responses to the predicted responses. According to a previous study[1], 500 voxels with the highest predictive power were adopted for this task. As shown in Table 2, our approach achieved the identification accuracies of 96.67% (116/120) and 90.83% (109/120) for the two participants, which were higher than that of the GWP model (92% and 72%) and GNet (90% and 73.33%). The correlation maps between measured and predicted responses for the two participants were shown in Fig. 5a. For most of the rows in the correlation maps, the elements on the diagonal were significantly larger than the others, indicating that our approach had excellent identification ability. In addition, we investigated the identification accuracies with 100, 500, 1000, and 2000 voxels. As shown in Fig. 5b, our approach reached the highest accuracies when 500 voxels were adopted.

# Discussion

In this work, we proposed a SCNN-based visual perception encoding model, which consisted of the SCNN feature extractor and voxel-wise response predictors. Different from the traditional Gabor and CNN-based approaches that used real-value computation, our model relied on the spike-driven SCNN that processed the visual information in a more brain-like manner. The proposed model achieved remarkable success in predicting the brain activity evoked by handwritten characters and natural images with a simple two-layer unsupervised SCNN using two publicly available datasets as the test bed. Moreover, we accomplished the image reconstruction and image identifications tasks with promising results based on our encoding models, indicating that our model can be applied to solve practical brain-reading problems.

Neural encoding models are usually proposed on the basis of different physiological hypotheses. Evaluating and comparing the encoding performance of the models can help us estimate the similarities between the hypotheses and the real physiological mechanism. The SCNN combines the computational rules of SNN and the network architecture of CNN, which have significant advantages on modeling the visual system. Motivated by this notion, we established the SCNN-based encoding model to predict the brain responses induced by different visual inputs. To extract meaningful visual features, we used a SCNN that consists of a DoG layer and a convolutional layer, which aimed to mimic information processing in the retina and visual cortex, respectively. In addition, we developed an efficient feature selection algorithm for encoding based on the neuronal population receptive field properties[33,34] of the fMRI voxels. Compared with other benchmark methods (Gabor and CNN based encoding models), our model showed better encoding performance on the experimental data, which demonstrated the superiority of SCNN in visual perception encoding.

A debate about whether the brain is supervised or unsupervised has always been discussed. Instead of the supervised CNNs, we used the unsupervised SCNN trained by STDP in our model. Our result suggests that the visual cortex, at least the early visual areas, is more likely to learn visual representations in an unsupervised manner. Additionally, the STDP-based SCNN has the following advantages on neural encoding. First, it is biologically plausible because STDP is a bioinspired learning rule. Second, it can deal with both labeled and unlabeled data. Third, it is more suitable for small sample datasets, such as fMRI.

The realization of neural decoding tasks is the basis for many brain-reading applications, such as the brain–computer interfaces (BCI)[35]. There are two types of decoding models: one is derived from the encoding model, and the other is directly built in an end-to-end fashion. The advantage of the former is that it can provide the voxel-level functional description while completing the decoding tasks[5]. However, recent breakthroughs in decoding were mainly made using the latter models[31,36,37]. In this work, we accomplished the downstream decoding tasks including image reconstruction and image identification based on the encoding model. The results show that our approach outperformed state-of-the-art methods on both of the decoding tasks. This finding further validates the effectiveness of our encoding model and

indicates that encoding-based approaches have promising application potential in solving decoding tasks.

In conclusion, this work provides a powerful tool for neural encoding. On the one hand, we combined the structure of CNNs and the calculation rules of SNNs to model the visual system and constructed voxel-wise encoding models based on the receptive field mechanism. On the other hand, we proved that our model can be used to implement practical decoding tasks, such as image reconstruction and image identification. We believe that the SCNN-based encoding models will provide crucial insights into the visual mechanism and contribute to solve BCI and computer vision tasks. Compared with deep-learning networks, the architectures of SNNs are usually shallower, which limits their abilities to extract complex and hierarchical visual features. Several recent studies have attempted to overcome this problem and made some progress[23,24,38]. The use of a deeper SCNN in our model will further improve the encoding performance and allow us to investigate the hierarchical structure of the visual cortex. Furthermore, we would like to extend the SNNs to the encoding tasks of other cognition functions (e.g., imagination and memory) in the future.

# Methods

**SCNN-based encoding model.** We proposed a SCNN-based encoding model to predict fMRI activities evoked by the input visual stimuli. The encoding model consisted of a SCNN feature extractor and voxel-wise regression models. In brief, for each input image, we first used the unsupervised SCNN to extract the stimulus features and then built linear regression models to project the SCNN features into fMRI responses. The architecture of the encoding model is shown in Fig. 1a.

*SCNN feature extractor.* We adopted a simple two-layer SCNN to extract stimuli features. The first layer refers to the Difference of Gaussians (DoG) layer, which is motivated to mimic the neural processing in the retinal ganglion cells[39,40]. In this layer, each input image is convolved by six DoG filters with zero padding. ON- and OFF-center DoG filters with size of $3 \times 3$, $7 \times 7$, and $13 \times 13$ and standard deviations of $(3/9, 6/9)$, $(7/9, 14/9)$, and $(13/9, 26/9)$ are used. The padding size is set to be 6. Subsequently, DoG features are converted into spike waves by intensity-to-latency encoding [41] with the length of 30. Specifically, we sorted DoG feature values greater than 50 in descending order and distributed them equally into 30 bins to generate the spike waves. Before passing to the next layer, the output spikes underwent max pooling operation with the window size of $2 \times 2$ and a stride of 2.

The second layer corresponds to the convolutional layer, which aims to mimic the information integration mechanism of the visual cortex. In this layer, we used 64 convolutional kernels that were made up with Integrate-and-Fire (IF) neurons to process the input spikes. The window size of the convolutional kernels is $5 \times 5$, and the padding size is 2. Each IF neuron gathers input spikes from its receptive field and emits a spike when its voltage reaches the threshold, which can be formulated as follows:

$$v_i\left(t\right)=v_i\left(t-1\right)+\sum_j w_{ij}\times s_j(t-1),$$

1

$$s_i\left(t\right)=1\mathrm{and}v_i\left(t\right)=0,if v_i\left(t\right)\geq v_{th},$$

2

where $v_i\left(t\right)$ denotes the voltage of $i_{th}$ IF neuron at time step t, $w_{ij}$ denotes the synaptic weight between $i_{th}$ neuron and $j_{th}$ input spikes within its receptive field, and $v_{th}$ denotes the firing threshold and is set to be 10. For each image, neurons are allowed to fire once at most. In addition, we adopted the inhibition mechanism in the convolutional layer. That is, one neuron, which is the neuron with the earliest spike time, is allowed at most to fire at each position in the feature maps. The synaptic weights are updated through STDP, which can be described as follows:

$$\Delta w_{ij}=\begin{cases}a^+\times w_{ij}\times\left(1-w_{ij}\right),if t_j-t_i\leq 0,\\ a^-\times w_{ij}\times\left(1-w_{ij}\right),if t_j-t_i>0,\end{cases}$$

3

where $\Delta w_{ij}$ denotes the weight modification; $a^+$ and $a^-$ denote the learning rates and are set to be 0.004 and − 0.003, respectively; and $t_i$ and $t_j$ denote the spike time of $i_{th}$ neuron and $j_{th}$ input spikes, respectively. We calculated the learning convergence defined in Kheradpisheh, et al. [23] using the following formula:

$$\mathrm{C}=\sum_i\sum_j w_{ij}\times(1-w_{ij})/\mathrm{N}$$

4

,

where N denotes the total number of synaptic weights. The training of the convolutional layer is stopped when $\mathrm{C}$ is less than 0.01. The implementation of SCNN was based on the SpykeTorch platform[42]. After training the SCNN, we set the firing threshold $v_{th}$ to be infinite and measured the voltage value at the last time step in each neuron as the SCNN feature of the visual stimuli. Given that the voltages in the convolutional neurons accumulated over time and has never been reset when $v_{th}$ was infinite, the final voltage values (SCNN feature) reflected the activations of SCNN to the visual stimuli.

*Responses prediction algorithm.* With the SCNN feature $\mathrm{F}\in\mathcal{R}^{64\times h\times w}$ obtained as mentioned above, we built the linear regression model for each voxel to predict the fMRI response $\mathrm{Y}$. To determine the optimal receptive field location for each voxel, we went through all the locations on the SCNN feature maps to fit

the regression model and performed 3-fold cross-validation on the training data. The expression and object function of the regression model are defined as follows:

$$y_v = w \times f_{ij} + \epsilon$$

5

,

$$\min_w \| w \times f_{ij} - y_v \|_2^2$$

6

,

where $y_v$ denotes the fMRI response of the voxel $v$, $w$ denotes the weight parameters in the regression model, and $f_{ij} \in \mathcal{R}^{64\times1}(i = 1,2,\ldots,h, j = 1,2,\ldots,w)$ denotes the feature vector at location $(i, j)$ of the SCNN feature maps. We quantified the regression accuracy with the coefficient of determination ( $R^2$) of the predicted and observed responses, and the feature location with the highest $R^2$ was selected as the receptive field location of each voxel. Finally, we retrained the regression model for each voxel on the whole training data based on the determined receptive field location.

**Downstream decoding tasks.** Based on the encoding models, we accomplished two downstream decoding tasks: image reconstruction and image identification. Given the observed fMRI response, the image reconstruction task aims to reconstruct the perceived image, and the image identification task aims to identify which image was seen. The detailed implementations of the two tasks are described as follows.

*Image reconstruction.* As shown in Fig. 1b, the implementation of the image reconstruction task relied on a large prior image set. First, we used the encoding model to produce the predicted fMRI responses for all the images in the prior image set. Second, we calculated the Pearson's correlation coefficients (PCCs) between the observed fMRI response and the predicted fMRI response of the prior images. Finally, we averaged the prior images whose predicted fMRI responses had the highest PCC with the observed one to obtain the reconstruction result.

*Image identification.* The implementation of the image identification task is shown in Fig. 1c. The images of the test data were inputted into the encoding model to produce the predicted fMRI responses. Then, the PCCs between the predicted fMRI responses and the observed fMRI response were calculated. The image whose predicted fMRI response was most correlated with the observed one was regarded as the image seen by the subject.

**fMRI datasets.** To validate the encoding model, we employed two publicly available datasets that have been widely used in previous studies[1,25,31,36,43]: handwritten character and natural image datasets. The

brief descriptions of the datasets are listed in the following, and the details can be found in the original literatures.

The handwritten character dataset contains the fMRI data from three participants as they viewed handwritten character images. A total of 360 images for 6 characters (B, R, A, I, N, and S) with the size of $56 \times 56$ were presented for each participant, which were taken from the TICH character dataset [44]. The fixation point was added to each image as a white square. During the experiment, each image was presented for 1 s (flashed at 2.5 Hz) followed by a 3 s black background, and 3 T fMRI data were acquired simultaneously (TR = 1.74 s, voxel size = $2 \times 2 \times 2\mathrm{mm}^3$). The voxel-level fMRI responses of visual areas V1 and V2 for each visual stimulus were estimated using general linear models[45]. We adopted the same train/test set split as in the original work[25], which contained 270 and 90 class-balanced examples, respectively.

The natural image dataset contains the fMRI data from two participants as they viewed natural images[1]. The experiment was divided into train and test stages. In the train stage, 1750 images were presented to the participants, and each image was shown for 1 s (flashed at 2 Hz) followed by a 3 s grey background. In the test stage, participants were shown with 120 images that were different from the images used in the train stage. 3T fMRI data were acquired simultaneously in both experiment stages (TR = 1 s, voxel size = $2 \times 2 \times 2.5\mathrm{mm}^3$). The voxel-level fMRI responses of visual areas V1, V2, and V3 for each visual stimulus were estimated. To reduce computational complexity, the natural images were down sampled from $500 \times 500$ to $128 \times 128$ pixels.

# Declarations

### Data Available

The Image-fMRI datasets used in this study are freely available through the website (http://sciencesanne.com/research/) and (https://crcns.org/datasets/vc/vim-1).

### Code Available

The code that supports the findings of this study is available from https://github.com/wang1239435478/Neural-encoding-with-unsupervised-spiking-convolutional-spiking-neural-networks.

### Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 82121003, 62036003, 62276051, 82072006), the National Science and Technology Innovation 2030 Major Program (2022ZD0208900), the Chengdu Science and Technology Project (2021-YF05-01211-SN), Medical-Engineering Cooperation Funds from University of Electronic Science and Technology of China

(ZYGX2021YGLH201), Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (No. ZYYCXTD-D-202003).

### Author contributions

Chong Wang conceived the project. Chong Wang and Wei Huang implemented the proposed encoding framework. Tao Liu and Yuting Wang preprocessed the image-fMRI datasets. Wei Sheng, Ting Zou and Yun-Shuang Fan contributed to the method comparison. Chong Wang wrote the original draft. Hongmei Yan and Rong Li reviewed and edited the manuscript. Huafu Chen, Hongmei Yan and Rong Li acquired the fundings.

### Competing interests

The authors declare no competing interests.

# References

1. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355, doi:10.1038/nature06713 (2008).
2. Güçlü, U. & van Gerven, M. A. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **35**, 10005–10014, doi:10.1523/jneurosci.5023-14.2015 (2015).
3. Nishimoto, S. *et al.* Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology : CB* **21**, 1641–1646, doi:10.1016/j.cub.2011.08.031 (2011).
4. Wen, H. *et al.* Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral cortex (New York, N.Y. : 1991)* **28**, 4136–4160, doi:10.1093/cercor/bhx268 (2018).
5. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410, doi:https://doi.org/10.1016/j.neuroimage.2010.07.073 (2011).
6. Wu, M. C. K., David, S. V. & Gallant, J. L. COMPLETE FUNCTIONAL CHARACTERIZATION OF SENSORY NEURONS BY SYSTEM IDENTIFICATION. *Annual Review of Neuroscience* **29**, 477–505, doi:10.1146/annurev.neuro.29.051605.113024 (2006).
7. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America. A, Optics and image science* **2**, 284–299, doi:10.1364/josaa.2.000284 (1985).
8. Jones, J. P. & Palmer, L. A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology* **58**, 1233–1258, doi:10.1152/jn.1987.58.6.1233 (1987).
9. Carandini, M. *et al.* Do we know what the early visual system does? *The Journal of neuroscience : the official journal of the Society for Neuroscience* **25**, 10577–10597, doi:10.1523/jneurosci.3726-05.2005 (2005).

10. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci* **17**, 401–412, doi:10.1016/j.tics.2013.06.007 (2013).
11. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology* **10**, e1003915, doi:10.1371/journal.pcbi.1003915 (2014).
12. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports* **6**, 27755, doi:10.1038/srep27755 (2016).
13. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience* **25**, 116–126, doi:10.1038/s41593-021-00962-x (2022).
14. Khosla, M., Ngo, G. H., Jamison, K., Kuceyeski, A. & Sabuncu, M. R. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances* **7**, eabe7547, doi:doi:10.1126/sciadv.abe7547 (2021).
15. Xu, Y. & Vaziri-Pashkam, M. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications* **12**, 2065, doi:10.1038/s41467-021-22244-7 (2021).
16. Maass, W. Networks of spiking neurons: The third generation of neural network models. *Neural Networks* **10**, 1659–1671, doi:https://doi.org/10.1016/S0893-6080(97)00011-7 (1997).
17. Gerstner, W., Kempter, R., van Hemmen, J. L. & Wagner, H. A neuronal learning rule for sub-millisecond temporal coding. *Nature* **383**, 76–78, doi:10.1038/383076a0 (1996).
18. Bi, G.-q. & Poo, M.-m. Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. *The Journal of Neuroscience* **18**, 10464, doi:10.1523/JNEUROSCI.18-24-10464.1998 (1998).
19. Huang, S. *et al.* Associative Hebbian synaptic plasticity in primate visual cortex. *Journal of Neuroscience* **34**, 7575–7579, doi:10.1523/JNEUROSCI.0983-14.2014 (2014).
20. McMahon, David B. T. & Leopold, David A. Stimulus Timing-Dependent Plasticity in High-Level Vision. *Current Biology* **22**, 332–337, doi:https://doi.org/10.1016/j.cub.2012.01.003 (2012).
21. Meliza, C. D. & Dan, Y. Receptive-Field Modification in Rat Visual Cortex Induced by Paired Visual Stimulation and Single-Cell Spiking. *Neuron* **49**, 183–189, doi:https://doi.org/10.1016/j.neuron.2005.12.009 (2006).
22. Diehl, P. & Cook, M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience* **9**, doi:10.3389/fncom.2015.00099 (2015).
23. Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J. & Masquelier, T. STDP-based spiking deep convolutional neural networks for object recognition. *Neural networks : the official journal of the International Neural Network Society* **99**, 56–67, doi:10.1016/j.neunet.2017.12.005 (2018).
24. Mozafari, M., Ganjtabesh, M., Nowzari-Dalini, A., Thorpe, S. J. & Masquelier, T. Bio-inspired digit recognition using reward-modulated spike-timing-dependent plasticity in deep convolutional networks. *Pattern Recognition* **94**, 87–95, doi:https://doi.org/10.1016/j.patcog.2019.05.015 (2019).

25. Schoenmakers, S., Barth, M., Heskes, T. & van Gerven, M. Linear reconstruction of perceived images from human brain activity. *Neuroimage* **83**, 951–961, doi:10.1016/j.neuroimage.2013.07.043 (2013).
26. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2014). <https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K>.
27. Seeliger, K. *et al.* End-to-end neural system identification with neural information flow. *PLoS computational biology* **17**, e1008558, doi:10.1371/journal.pcbi.1008558 (2021).
28. Miyawaki, Y. *et al.* Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders. *Neuron* **60**, 915–929, doi:https://doi.org/10.1016/j.neuron.2008.11.004 (2008).
29. Du, C., Du, C. & He, H. Sharing deep generative representation for perceived image reconstruction from human brain activity. *Proceedings of the International Joint Conference on Neural Networks* **2017-May**, 1049–1056, doi:10.1109/IJCNN.2017.7965968 (2017).
30. Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y. & van Gerven, M. A. J. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage* **181**, 775–785, doi:https://doi.org/10.1016/j.neuroimage.2018.07.043 (2018).
31. Ren, Z. *et al.* Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage* **228**, 117602, doi:https://doi.org/10.1016/j.neuroimage.2020.117602 (2021).
32. Zhou, W., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**, 600–612, doi:10.1109/TIP.2003.819861 (2004).
33. Victor, J. D., Purpura, K., Katz, E. & Mao, B. Population encoding of spatial frequency, orientation, and color in macaque V1. *Journal of neurophysiology* **72**, 2151–2166, doi:10.1152/jn.1994.72.5.2151 (1994).
34. Dumoulin, S. O. & Wandell, B. A. Population receptive field estimates in human visual cortex. *NeuroImage* **39**, 647–660, doi:https://doi.org/10.1016/j.neuroimage.2007.09.034 (2008).
35. Gao, X., Wang, Y., Chen, X. & Gao, S. Interface, interaction, and intelligence in generalized brain-computer interfaces. *Trends Cogn Sci* **25**, 671–684, doi:10.1016/j.tics.2021.04.003 (2021).
36. Du, C., Du, C., Huang, L. & He, H. Reconstructing Perceived Images From Human Brain Activities With Bayesian Deep Multiview Learning. *IEEE Transactions on Neural Networks and Learning Systems* **30**, 2310–2323, doi:10.1109/TNNLS.2018.2882456 (2019).
37. Wang, C. *et al.* Reconstructing Rapid Natural Vision with fMRI-Conditional Video Generative Adversarial Network. *Cerebral Cortex*, doi:10.1093/cercor/bhab498 (2022).
38. Wu, Y., Deng, L., Li, G., Zhu, J. & Shi, L. Spatio-Temporal Backpropagation for Training High-Performance Spiking Neural Networks. *Front Neurosci* **12**, 331, doi:10.3389/fnins.2018.00331 (2018).
39. Enroth-Cugell, C. & Robson, J. G. The contrast sensitivity of retinal ganglion cells of the cat. *J Physiol* **187**, 517–552, doi:10.1113/jphysiol.1966.sp008107 (1966).

40. McMahon, M. J., Packer, O. S. & Dacey, D. M. The classical receptive field surround of primate parasol ganglion cells is mediated primarily by a non-GABAergic pathway. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **24**, 3736–3745, doi:10.1523/jneurosci.5252-03.2004 (2004).
41. Gautrais, J. & Thorpe, S. Rate coding versus temporal order coding: a theoretical approach. *Biosystems* **48**, 57–65, doi:https://doi.org/10.1016/S0303-2647(98)00050-1 (1998).
42. Mozafari, M., Ganjtabesh, M., Nowzari-Dalini, A. & Masquelier, T. SpykeTorch: Efficient Simulation of Convolutional Spiking Neural Networks With at Most One Spike per Neuron. *Frontiers in Neuroscience* **13**, doi:10.3389/fnins.2019.00625 (2019).
43. Du, C., Du, C., Huang, L. & He, H. in *National Conference on Artificial Intelligence.*
44. Van der Maaten, L. A new benchmark dataset for handwritten character recognition. *Tilburg University*, 2–5 (2009).
45. Friston, K. J. *et al.* Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping* **2**, 189–210, doi:https://doi.org/10.1002/hbm.460020402 (1994).

# Figures

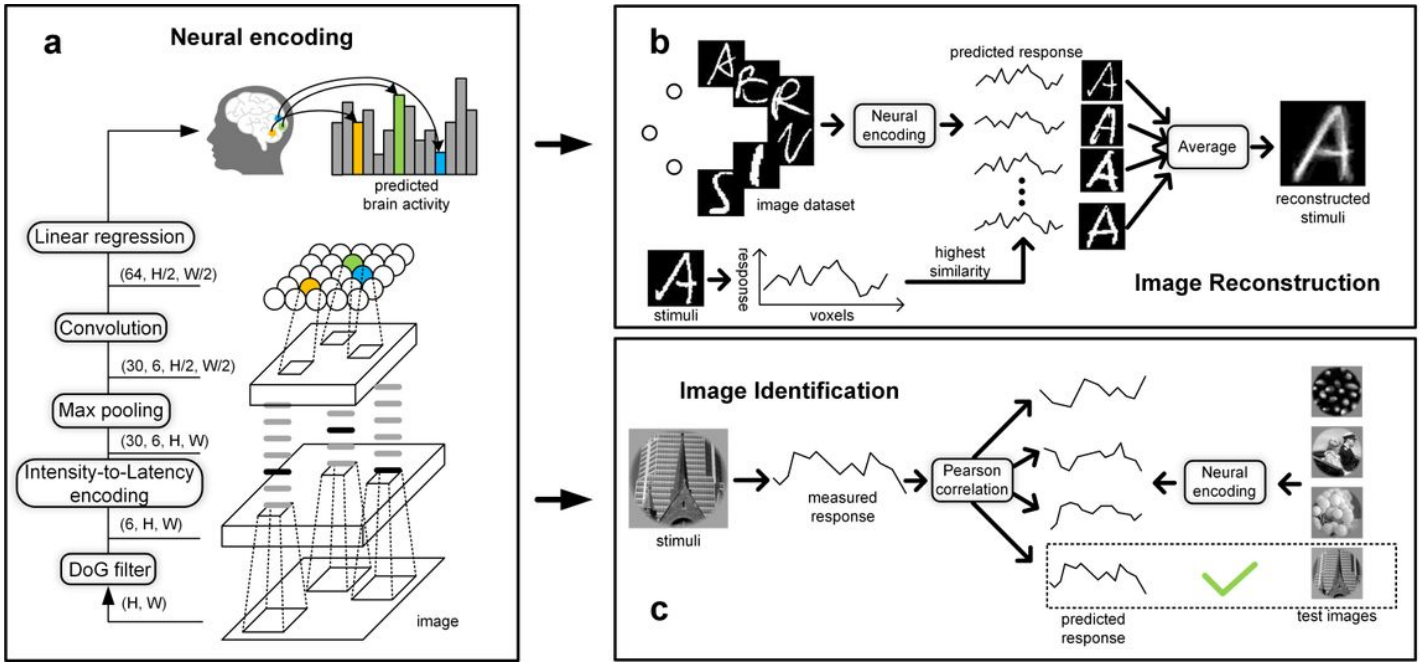
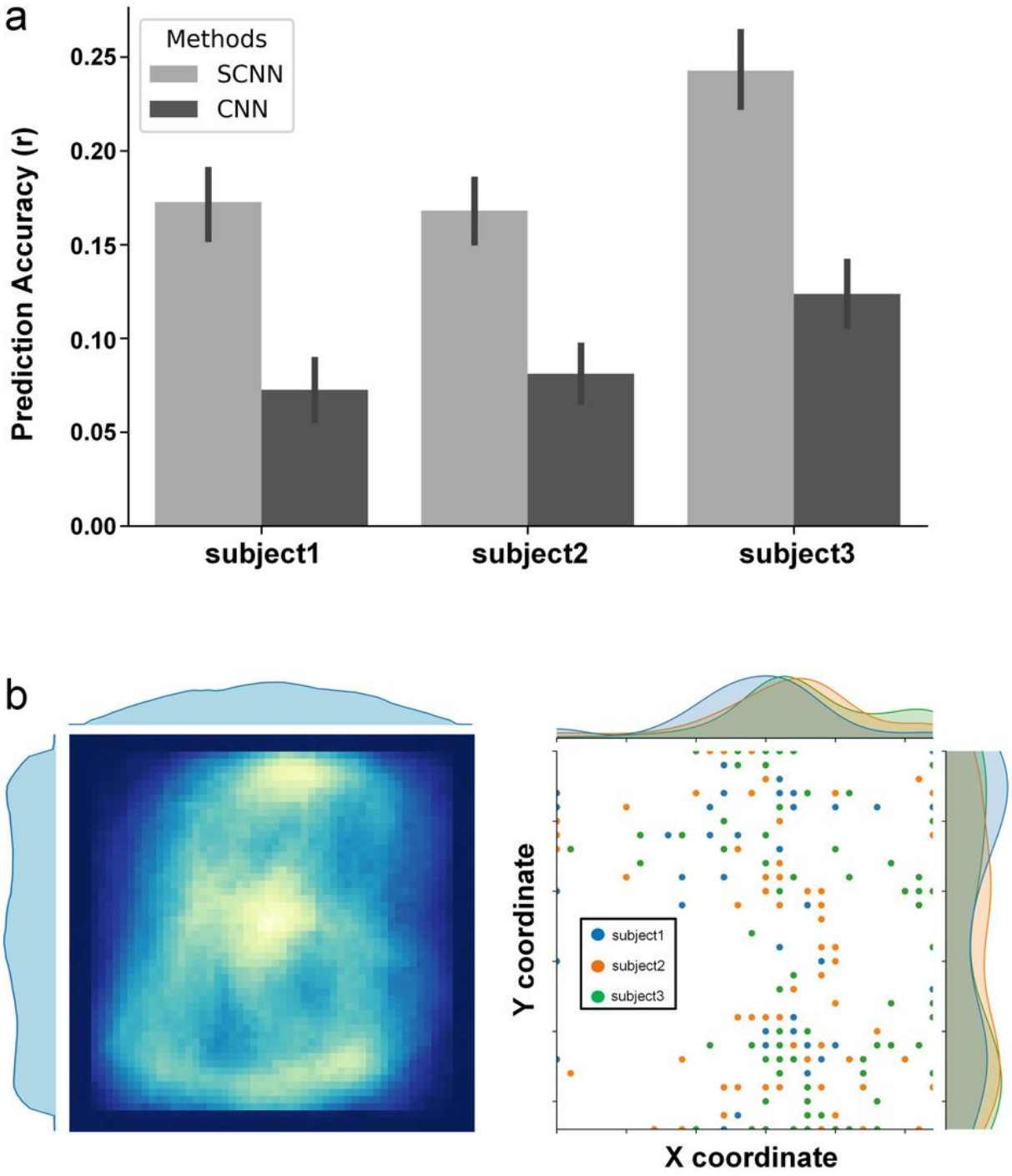



**Figure 1**

**The flow chart of the encoding and decoding processes. a** The illustration of the encoding model. The proposed model uses a two-layer SCNN to extract visual features of the input images, and uses linear regression models to predict the fMRI responses for each voxel. **b** The diagram for the image reconstruction task, which aims to reconstruct the perceived images from the brain activity. **c** The

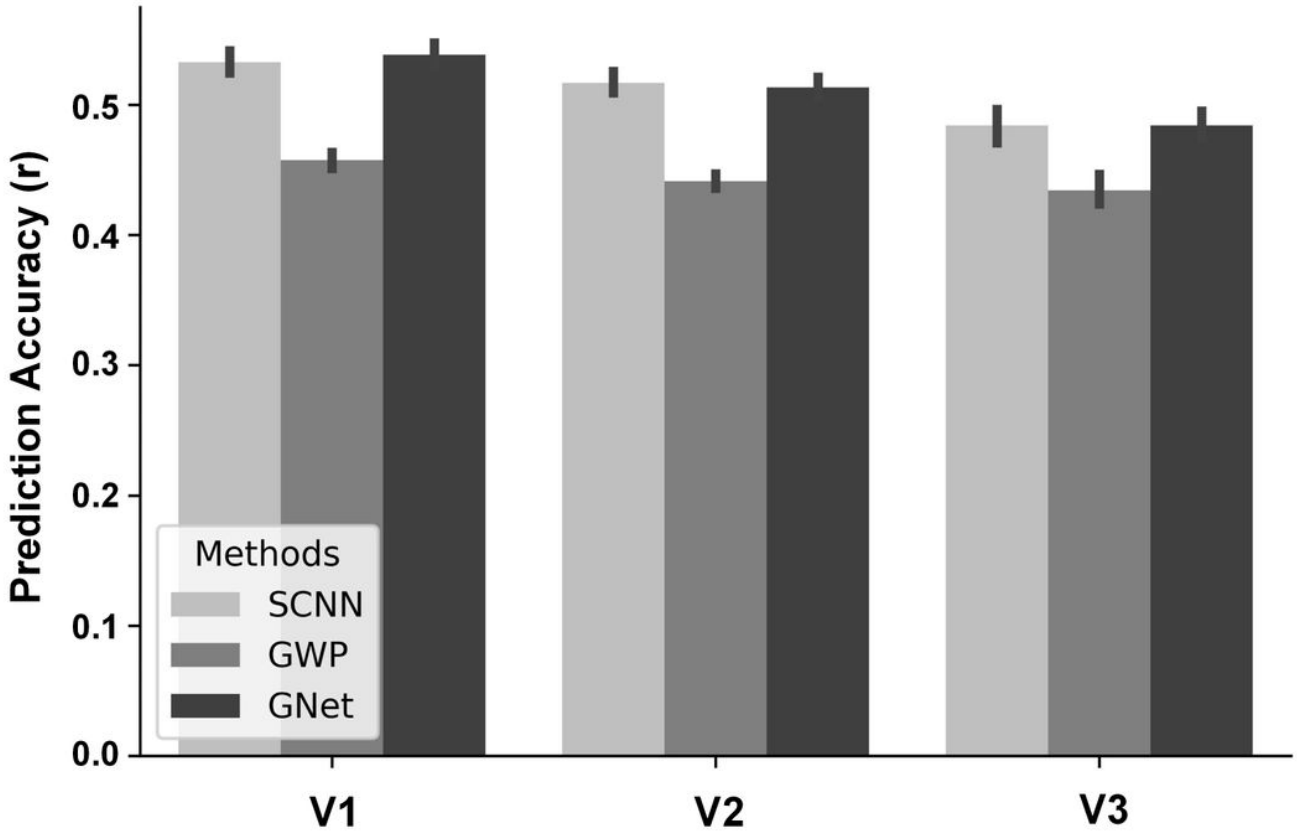
diagram for the image identification task, which aims to identify which image is perceived based on the fMRI responses.




## Figure 2

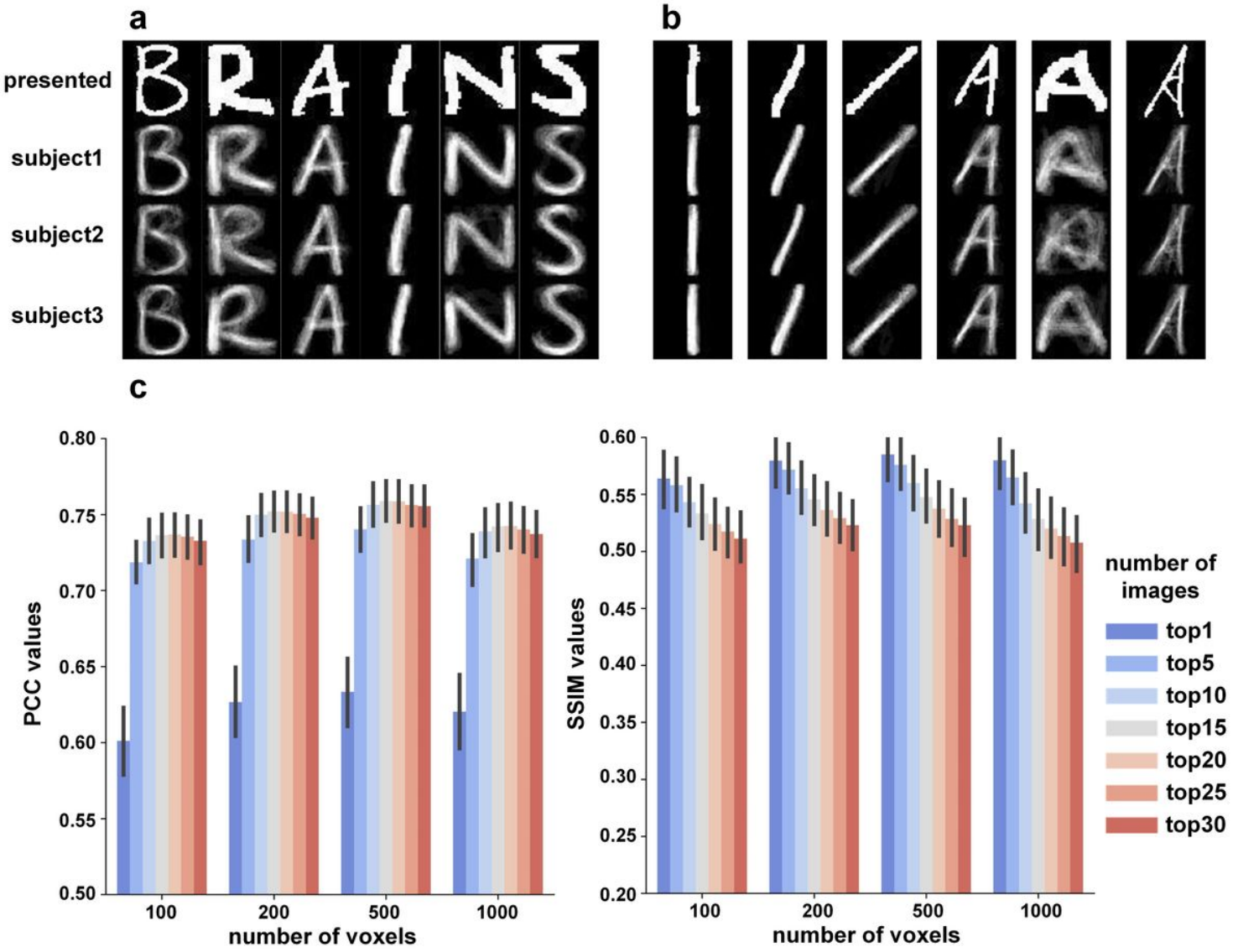
**The encoding results on handwritten character dataset. a** The encoding performance (mean ± SEM) of SCNN and CNN-based models on the selected voxels. The accuracies of SCNN are significantly higher

than those of CNN for all the three subjects ($p < 10^{-13}$, one-tailed two sample t-test). **b** The mean intensities of the images in the train set (left) and receptive field locations of the 100 most predictable voxels (right).




## Figure 3

**The encoding performance of different encoding models on natural image dataset.** The bar charts show the prediction accuracies of different visual areas averaged across each predictable voxels and each subject, and the error bars show the standard error of the mean (SEM). For each visual areas, the accuracies of SCNN were significantly higher than those of GWP ($p < 10^{-13}$, one-tailed two sample t-test) and have no significantly difference with GNet ($p > 0.37$, two-tailed two sample t-test).




## Figure 4

**Reconstruction results on handwritten character dataset. a** The reconstructions of different characters (B, R, A, I, N, and S). **b** The reconstructions of the same character with different writing styles. The images in the first row are the presented images (ground truth), and the images in the second to fourth rows are the reconstruction results of the 3 participants. **c** The reconstruction performance (PCC and SSIM) of different hyper-parameters (number of selected voxels and number of averaged images) on handwritten character dataset.

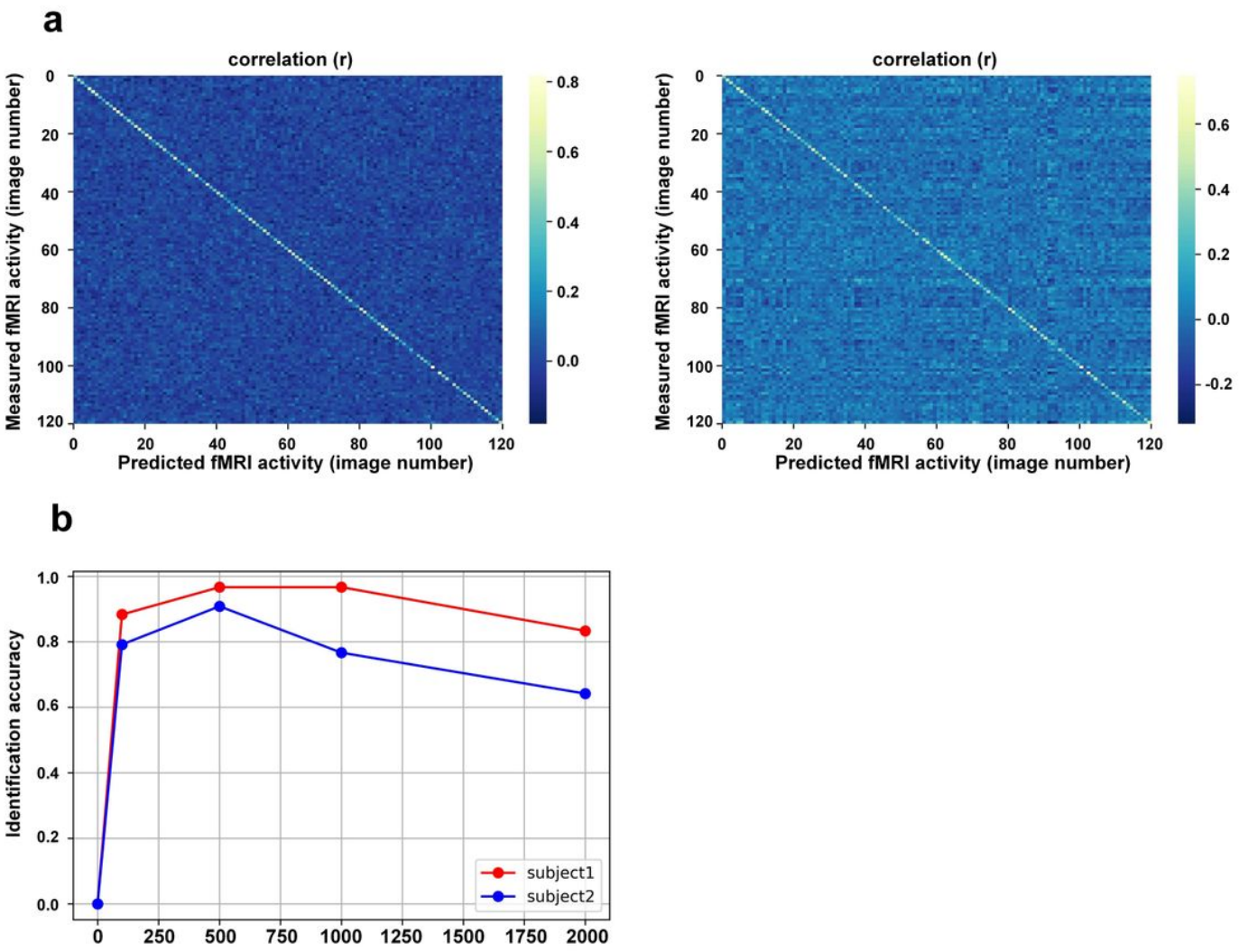



**Figure 5**

**Image identification results on natural image dataset. a** The correlation maps of the measured and predicted fMRI responses to test set images for the two participants. The element at the $m^{th}$ column and $n^{th}$ row represents the correlation between the measured fMRI response for the $m^{th}$ image and the predicted fMRI response for the $n^{th}$ image. **b**The identification accuracies with different numbers of voxels for the two participants in natural image dataset.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf