frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 22 June 2018
doi: 10.3389/fnhum.2018.00242

Constraint-Free Natural Image Reconstruction From fMRI Signals Based on Convolutional Neural Network

Chi Zhang[1], *Kai Qiao*[1], *Linyuan Wang*[1], *Li Tong*[1], *Ying Zeng*[1,2] *and Bin Yan*[1]*

[1] *National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China,* [2] *Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China*

In recent years, research on decoding brain activity based on functional magnetic resonance imaging (fMRI) has made eye-catching achievements. However, constraint-free natural image reconstruction from brain activity remains a challenge, as specifying brain activity for all possible images is impractical. The problem was often simplified by using semantic prior information or just reconstructing simple images, including digitals and letters. Without semantic prior information, we present a novel method to reconstruct natural images from the fMRI signals of human visual cortex based on the computation model of convolutional neural network (CNN). First, we extracted the unit output of viewed natural images in each layer of a pre-trained CNN as CNN features. Second, we transformed image reconstruction from fMRI signals into the problem of CNN feature visualization by training a sparse linear regression to map from the fMRI patterns to CNN features. By iteratively optimization to find the matched image, whose CNN unit features become most similar to those predicted from the brain activity, we finally achieved the promising results for the challenging constraint-free natural image reconstruction. The semantic prior information of the stimuli was not used when training decoding model, and any category of images (not constraint by the training set) could be reconstructed theoretically. We found that the reconstructed images resembled the natural stimuli, especially in position and shape. The experimental results suggest that hierarchical visual features may be an effective tool to express the human visual processing.

Keywords: image reconstruction, functional magnetic resonance imaging, convolutional neural network, visual representation, brain decoding

OPEN ACCESS

Edited by:
Jay Hegdé,
Augusta University, United States

Reviewed by:
Gopikrishna Deshpande,
Auburn University, United States
Hidenao Fukuyama,
Kyoto University, Japan

****Correspondence:***
Bin Yan
ybspace@hotmail.com

Received: *03 February 2018*
Accepted: *28 May 2018*
Published: *22 June 2018*

Citation:
Zhang C, Qiao K, Wang L, Tong L, Zeng Y and Yan B (2018) Constraint-Free Natural Image Reconstruction From fMRI Signals Based on Convolutional Neural Network.
Front. Hum. Neurosci. 12:242.
doi: 10.3389/fnhum.2018.00242

INTRODUCTION

Functional magnetic resonance imaging (fMRI) has become an effective tool for decoding brain activity, especially in visual decoding. A large number of studies have implemented the classification of stimulus categories (Mitchell et al., 2008; Huth et al., 2012), memories (Postle, 2015), imagination (Reddy et al., 2010), and even dreams (Horikawa et al., 2013) by multi-voxel pattern analysis (MVPA) (Zafar et al., 2015). More precisely, encoding model has been built to identify stimulus (Kay et al., 2008). Very few studies focused on visual image reconstruction. The goal of reconstruction is to produce a literal picture of the stimulus image. Visual image reconstruction is a more challenging problem because it needs much more decoded information than classification or identification, especially for natural images containing infinitely variable complex information.

To simplify the problem of stimulus image reconstruction, most studies focused on the reconstruction of simple images. Thirion et al. (2006) first implemented image reconstruction based on fMRI. They estimated the response model of each voxel in the retinotopic mapping experiment and reconstructed the simple images composed of quickly rotating Gabor filters in the passive viewing experiment and imagery experiment for the same subject. Miyawaki et al. (2008) realized the reconstruction of simple letters and graphics (10 × 10 resolution) by solving the linear mapping model from the voxels of visual cortex to each pixel of image. Schoenmakers et al. (2013) introduced the idea of sparse learning and the forward linear Gauss model to reconstruct the handwritten English letter "BRAINS" from the fMRI signals of the visual cortex. They further improved the results of letter reconstruction by introducing the Gauss hybrid model (Schoenmakers et al., 2014a,b). Yargholi and Hossein-Zadeh (2015, 2016) used the Gauss Network to reconstruct six and nine digital handwritten numerals, but it is more like a problem of classification in essence. Naselaris et al. (2009) first implemented the reconstruction of natural images using a priori information and a combination of structural coding and semantic coding models. However, it is essentially an image recognition problem in a limited natural image library. On this basis, they realized the reconstruction of video via image reconstruction frame by frame (Nishimoto et al., 2011).

At the same time, deep neural network (DNN) has become the focus of scholars in recent years due to its strong capability of feature representation. Deep learning has achieved a breakthrough in image detection/classification (Krizhevsky et al., 2012; Denton et al., 2015), speech recognition (Deng et al., 2013) and natural language processing (Blunsom et al., 2014; Le and Mikolov, 2014). More and more research have applied DNN to fMRI visual decoding (Yamins and DiCarlo, 2016). Agrawal et al. (2014) first encoded fMRI signals using the features extracted from images by convolutional neural network (CNN). Güçlü et al. used a DNN tuned for object categorization to probe neural responses to naturalistic stimuli. The result showed an explicit gradient for feature complexity existed from early visual areas toward the ventral (Güçlü and van Gerven, 2015a) and dorsal (Güçlü and van Gerven, 2015b) streams of the human brain. Cichy et al. (2016) compared temporal (magnetoencephalography, MEG) and spatial (fMRI) brain visual representations with representations in the DNN tuned to the statistics of real-world visual recognition. The results showed that the DNN captured the visual perception process of the human brain in both time and space in the ventral and dorsal visual pathways of the human brain. Horikawa and Kamitani (2017a) proposed a generic decoding model based on hierarchical visual features generated by DNN. They found that hierarchical visual features could be predicted from fMRI patterns and used them to identify seen/imagined object categories from a set of computed features for numerous object images. Furthermore, they found that the features decoded from the dream fMRI data had a strong positive correlation with the intermediate and advanced DNN layer features of the dreamed objects (Horikawa and Kamitani, 2017b). Du et al. (2017) achieved better performance in simple images reconstruction through deep generation networks, but this method still has some problems with natural image reconstruction. Using the convolution kernels of the first layer of CNN, Wen et al. (2017) implemented the reconstruction of dynamic video frame by frame. However, the results still had a gap with natural images, although the position information was restored well. In a word, all these studies suggested that DNN could help in providing more detailed interpretation of human brain visual information. Constraint-free natural images may be reconstructed well due to the efficient feature representation of DNN.

Recently, Mahendran and Vedaldi (2015, 2016) proposed a method about the input image generation for each CNN layer feature. Inspired by the research, this paper presents a novel visual image reconstruction method for natural images based on fMRI (**Figure 1**). By training the decoders that predict the CNN features of natural stimuli from fMRI activity patterns, we transformed image reconstruction from fMRI signals into the problem of CNN feature visualization. Then, iteratively optimization was performed to find the matched image whose CNN unit features became most similar to those predicted from the brain activity. Finally, the matched image was taken as the reconstruction result from the brain activity. By analyzing the experimental results, we verified the effectiveness of the method and the homology between human and computer visions.

MATERIALS AND METHODS

Experimental Data

The data used in this paper were the same as Kay et al. (2008), downloaded from an online data sharing database (http://crcns.org/data-sets/vc/vim-1). The data consisted of the blood-oxygen level dependent (BOLD) activities of two human subjects (S1 and S2) acquired using a 4T INOVA MR scanner (Varian, Inc., Palo Alto, CA, USA). EPI (echo-planar imaging) scan was performed to acquire functional images covering occipital cortex (repetition time (TR), 1,000 ms; echo time (TE), 28 ms; flip angle, 20°; field of view (FOV), 128 × 128 mm^2; slice thickness, 2.25 mm; slice gap, 0.25 mm; matrix size, 64 × 64; spatial resolution, 2 × 2 × 2.5 mm^3).

The dataset is divided into two sets: training set and validation set. In the training phase, the subjects viewed 1,750 grayscale natural images (20° × 20°) randomly selected from a database. Images were flashed at 200 ms intervals for 1 s followed by 3 s of gray background in successive 4 s trials during which subjects were fixated on a central white square (0.2° × 0.2°). During the validation phase, the subjects viewed 120 novel natural images presented in the same way as the training phase. Training images were presented 2 times each, and test images were presented 13 times each. Training and validation data were acquired in the same scan sessions.

First, functional images were manually co-registered to correct differences in head positioning across different sessions. Then, automated motion correction and slice timing were applied to

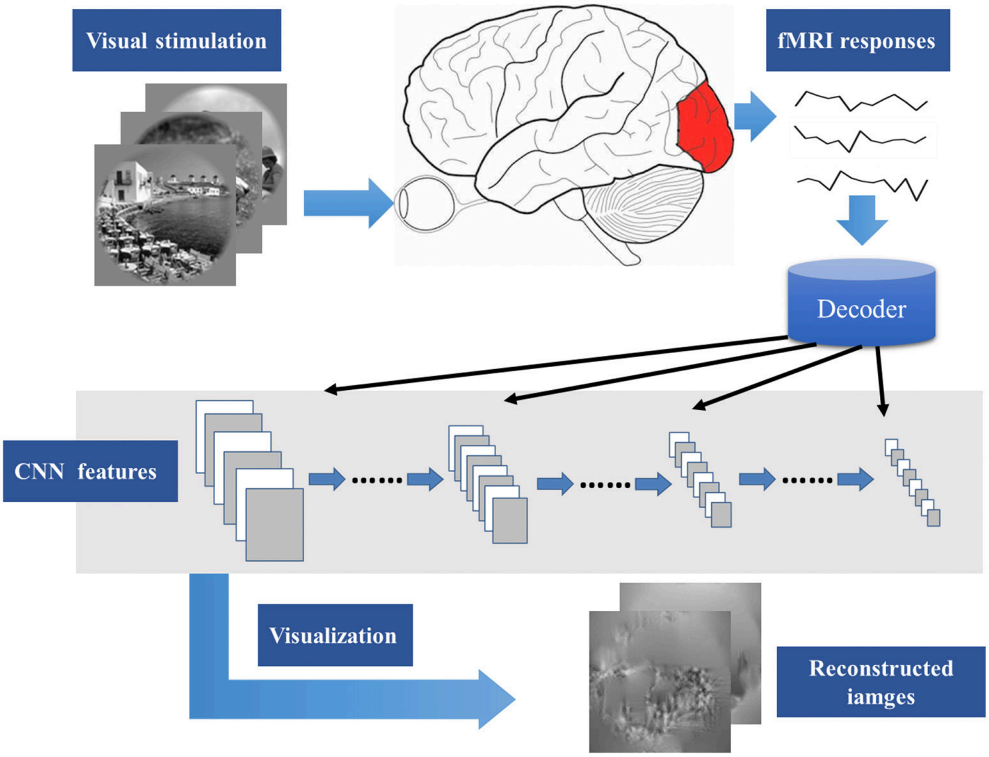

FIGURE 1 | Main process of visual image reconstruction. When the subject was seeing natural stimuli, fMRI responses were acquired through a MRI scanner. Then, the CNN features of natural stimuli were predicted by the decoder trained on the training set. The predicted CNN features were visualized by iteratively optimization to find the matched image whose CNN unit features became most similar to those predicted from the brain activity. Finally, the matched image was taken as the reconstruction result from the brain activity.

the data acquired within the same session by SPM (http://www.fil.ion.ucl.ac.uk/spm) software.

Extracting Hierarchical Visual Features Based on CNN

We used a deep CNN (Caffe–Alex [caffe]) (Jia et al., 2014), which closely reproduced the network by Krizhevsky et al. (2012) to extract hierarchical visual features from the stimuli. **Table 1** details the structure of the Caffe–Alex model. It is composed of the following computational building blocks: linear convolution, rectified linear unit (ReLU) gating, spatial max-pooling, and group normalization. This CNN was trained to achieve the best performance of object recognition in Large Scale Visual Recognition Challenge 2012.

This model can be concisely divided into eight layers: the first five are convolutional layers (consist of 96, 256, 384, 384, and 256 kernels), and the last three layers are fully connected for object classification (consist of 4,096, 4,096, and 1,000 artificial neurons). Each convolutional layer consists of some or all of the following four stages: linear convolution, ReLU gating, spatial max-pooling, and group normalization. Layers 6 and 7 are fully connected networks, and layer 8 uses a softmax function to output a vector of probabilities by which the input image is classified into individual categories.

For each image inputted to the CNN, the output of each layer was extracted to form the image hierarchy features. The dimensions of each layer features are shown in **Table 1**. We used the matconvnet toolbox (Vedaldi and Lenc, 2015) for implementing CNNs.

Decoding fMRI Signals to CNN Features

Using the training images, we estimated multivariate regression models to predict the feature maps of CNN layers based on distributed cortical fMRI signals. For each layer, a linear model was defined to map the distributed fMRI signals to the output features of artificial neurons in the CNN. For a specific feature of a particular CNN layer, it is expressed as Equation (1):

$$y = \mathbf{X}w, \tag{1}$$

where, y stands for the CNN features of training images, which is an m-by-1 matrix, where m is the number of training images. $\boldsymbol{X}$ stands for the observed fMRI signals within the visual cortex, which is an m-by-(n+1) matrix, where m is the number of training images, and n is the number of voxels. The last column of $\boldsymbol{X}$ is a constant vector with all elements equal to 1. w is the unknown weighting vector to solve. It is an (n+1)-by-1 matrix.

As the number of training samples m is far less than the number of voxels in visual field n, the problem is actually the

TABLE 1 | Structure of Caffe–Alex model.

Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Name	conv1	relu1	mpool1	norm1	conv2	relu2	mpool2	norm2	conv3	relu3	conv4	relu4	conv5	relu5	mpool5	fc6	relu6	fc7	relu7	fc8
Type	cnv	relu	mpool	nrm	cnv	relu	mpool	nrm	cnv	relu	cnv	relu	cnv	relu	mpool	cnv	relu	cnv	relu	cnv
Channels	96	96	96	96	256	256	256	256	384	384	384	384	256	256	256	4,096	4,096	4,096	4,096	1,000

solution of ill-posed equation, and no unique solution can be found. In addition, several theoretical studies suggest that a sparse coding scheme is used to represent natural images in primary visual cortex (Vinje and Gallant, 2000; Cox and Savoy, 2003). It means that only a small number of active neurons are for a special stimulus. By contrast, only a small number of visual stimuli can make a neuron active. As a proxy of neural activities, expecting that the responses of neurons can also reflect the sparse property is reasonable. Thus, w should be sparse to be more in line with visual characteristics.

Based on the above assumption, the major problem of constructing is how to solve a sparse representation problem. Traditional sparse recovery is formulated as a general NP-Hard problem as follows:

$$\min_{w} \|w\|_0 \; subject\ to\ \mathbf{X}w = y \tag{2}$$

Two approximate solutions could be used to solve the problem. One is transforming the NP-Hard L0 optimization problem into the L1 optimization problem. Donoho et al. showed that for some measurement matrix **X**, this NP-Hard problem is equivalent to its relaxation (Donoho and Stark, 1989):

$$\min_{w} \|w\|_1 \; subject\ to\ \mathbf{X}w = y. \tag{3}$$

L1-minimization method provides uniform guarantees for sparse recovery. If the measurement matrix satisfies the restricted isometry property (RIP) condition, it works correctly for all sparse signals. In this paper, we used YAll1 (Yang and Zhang, 2011) to solve the L1 optimization problem.

An alternate approach for sparse recovery problem is greedy algorithm. Greedy algorithms are quite fast by computing the support of the sparse signal iteratively, although it lacks the strong guarantees which L1-minimization provides. Considering that decoding model must be estimated for each CNN feature, the approximation method should be fast enough and simple to decrease the time cost. Therefore, we focused more on greedy algorithms to investigate the sparseness of decoding model. In this paper, we selected regularized orthogonal matching pursuit (ROMP) (Needell and Vershynin, 2009, 2010) to solve the decoding model. Finally, we compared both YAll1 and ROMP and selected ROMP as the solution for the decoding model.

Reconstructing Image From CNN Features

In a recent study, Mahendran et al. proposed a method to reconstruct original images from CNN features by gradient descent optimization (Mahendran and Vedaldi, 2015) to better understand deep image representations. This paper used the method to reconstruct the image from the decoded CNN features. We provided representation function $\Phi : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{D}$ (represents the process of the extracting CNN features of a layer) and decoded the CNN features of one layer $\Phi_0 = \Phi(x_0)$. The image reconstruction aims at finding the image

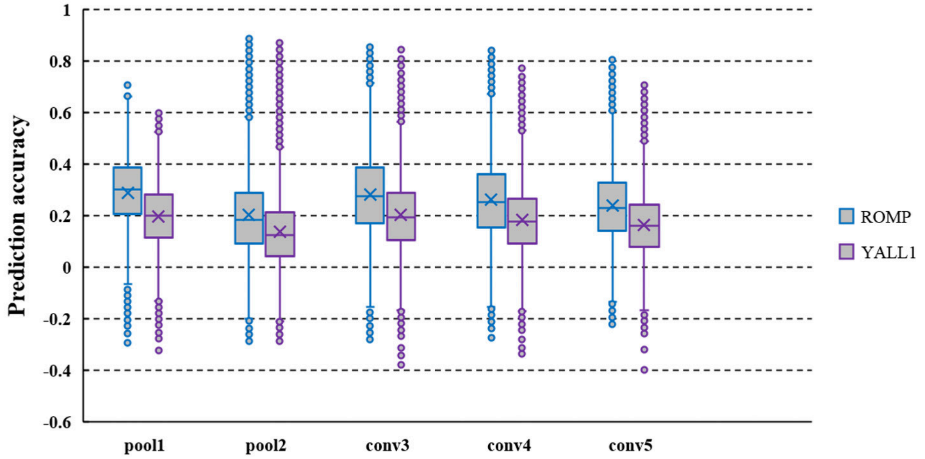

FIGURE 2 | Prediction accuracy of the pool1, pool2, conv3, conv4, and conv5 layer features decoded from the fMRI data of S1. CNN feature prediction accuracy are defined as the Pearson's correlation coefficient (r) between their actual and predicted feature values on the validation set. The average prediction accuracy of ROMP was significantly higher than that of YALL1 on t-statistics at a significance level of 10^{-5}.

$x \in \mathbb{R}^{H\times W\times C}$ that minimizes the following objective:

$$x^* = \underset{x\in\mathbb{R}^{H\times W\times C}}{\arg\min}\ \ell(\Phi(x), \Phi_0) + \lambda\,\Re(x), \tag{4}$$

where the loss ℓ compares the image representation $\Phi(x)$ to the target one Φ_0, and $\Re : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}$ stands for regularized constraint item. We used the Euclidean distance as the loss function and the regularized constraint item constants of two regularizers. To encourage images to stay within a target interval instead of diverging and to consist of piece-wise constant patches, the first one is simply the norm $\Re_\alpha(x) = \|x\|_\alpha^\alpha$ ($\alpha = 6$ is used in the experiments and$x$ is the vectorized and meansubtracted image) and the second richer regularizer is the total variation (TV). In addition, extended gradient descent used momentum (Krizhevsky et al., 2012) to solve (4) more effectively.

Quantification of Model Performance

To quantify how well the voxel responses predicted CNN features, we defined CNN feature prediction accuracy as the Pearson's correlation coefficient (r) between their actual and

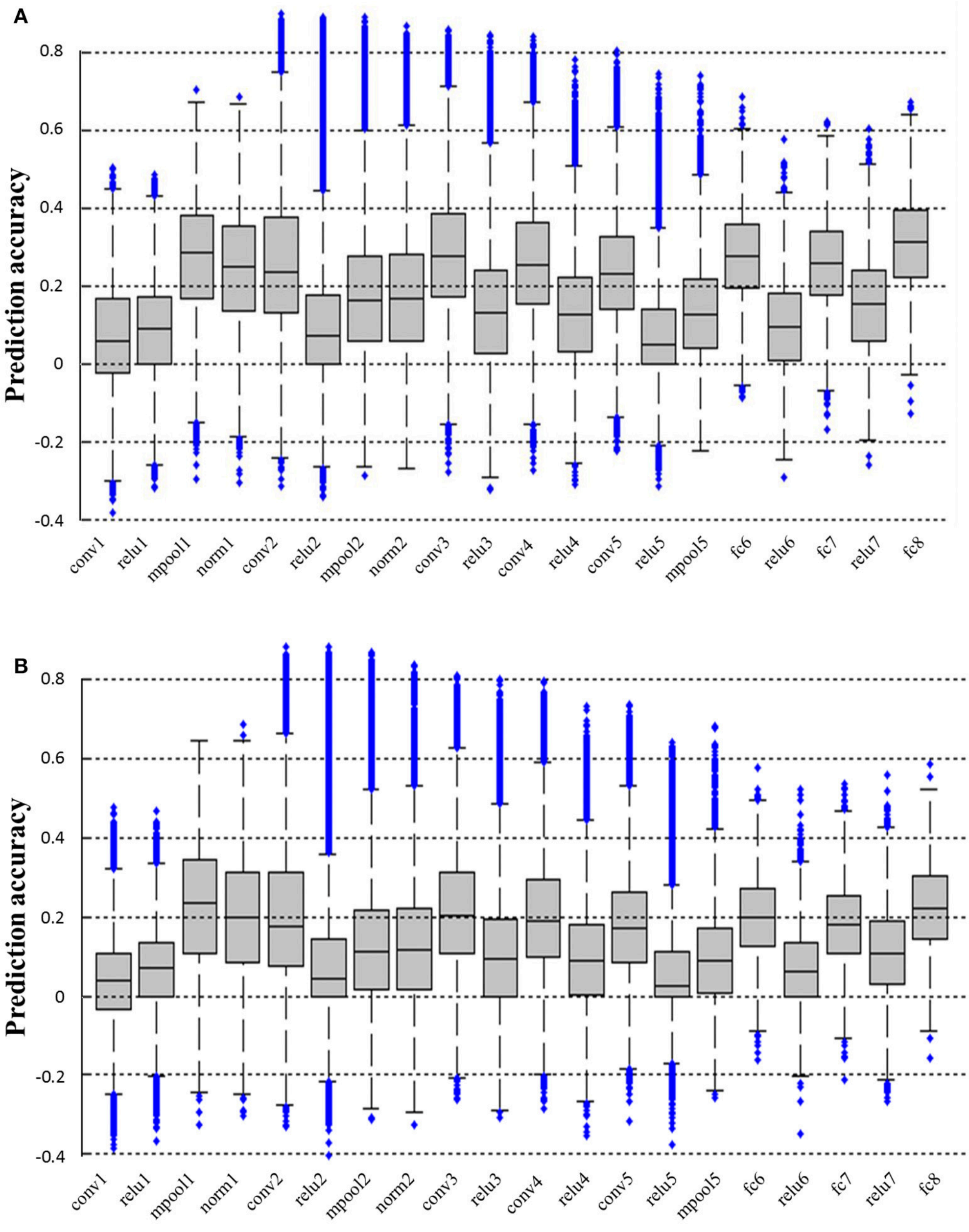

FIGURE 3 | Prediction accuracy of all layers of CNN. **(A,B)** Show the prediction accuracy of all layers based on fMRI signals of S1 and S2, respectively. All the prediction accuracy levels are significantly higher than chance (p < 0.01, T-test). The prediction accuracy levels of pool1, conv2, conv3, conv4, conv5, fc6, fc7, and fc8 layers are all significantly higher than those of conv1, relu1, relu2, relu3, relu4, relu5, relu6, and relu7 layers for both subjects (p < 0.001, T-test).

predicted feature values on the validation set. To achieve better image reconstruction performance, all the decoded features of one layer for image reconstruction should be similar to the features extracted from the actual natural image as far as possible. For a layer of CNN features, the mean *r* was used to express its prediction accuracy.

To solve (2) more precisely and achieve better image reconstruction, we compared ROMP and YALL1 with prediction accuracy. Moreover, we compared the prediction accuracy of different layers of CNN to find a most suitable layer to decode fMRI signals into and reconstruct the original image from. Finally, considering the prediction accuracy of each CNN layer and the characteristics of the image reconstruction method (with the same prediction accuracy, more low-level features, and better reconstruction performance), we selected pool1 layer as the image representation Φ (see section Reconstructing Image From CNN Features for more details) to reconstruct the original image from the voxel response.

Given the sparsity of the decoding model, the decoding process included voxel selection. During the decoding of fMRI signals into the features of each layer, we selected the 300 most frequently utilized voxels as significant voxels and defined the contribution of each visual area (V1, V2, V3, and V4) as the proportions of each visual area in the significant voxels. By

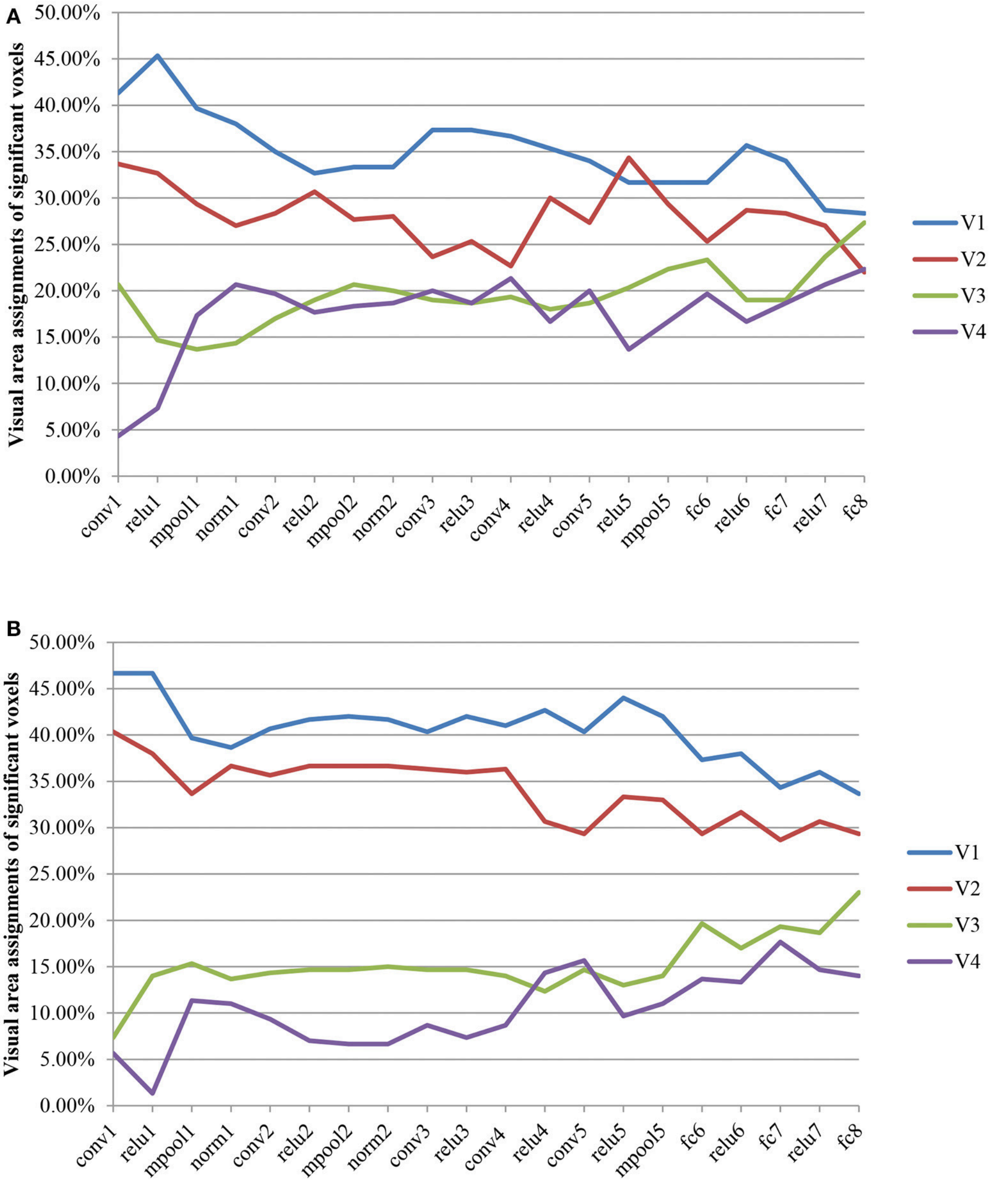

FIGURE 4 | Visual area assignments of significant voxels. **(A,B)** Show the result of S1 an S2, respectively. The total numbers of voxels in V1, V2, V3, and V4 are 1294/1399, 2083/1890, 1790/1772, and 484/556 (S1/S2). For each subject, 300 voxels with the highest frequency selected for decoding each layer feature served as significant voxels.

analyzing the segmentation of significant voxels, the hierarchical structure similarity between CNN and the visual cortex can be also verified.

Both objective and subjective assessment methods were used to assess the accuracy of the reconstructions. For the objective assessment, we calculated the weighted complex wavelet structural similarity metric (CWSSIM) to assess the accuracy of the reconstructions (Brooks et al., 2008). The metric used the coefficients of a complex wavelet decomposition of two images to compute a single number that described the degree of structural similarity between the two images. For the subjective assessment, we performed a behavioral experiment, in which a group of 13 raters (4 females and 9 males, aged from 22 to 38) were presented with one original image from the validation set and had to choose a similar one between the real and one randomly chosen different reconstruction taken from the same validation

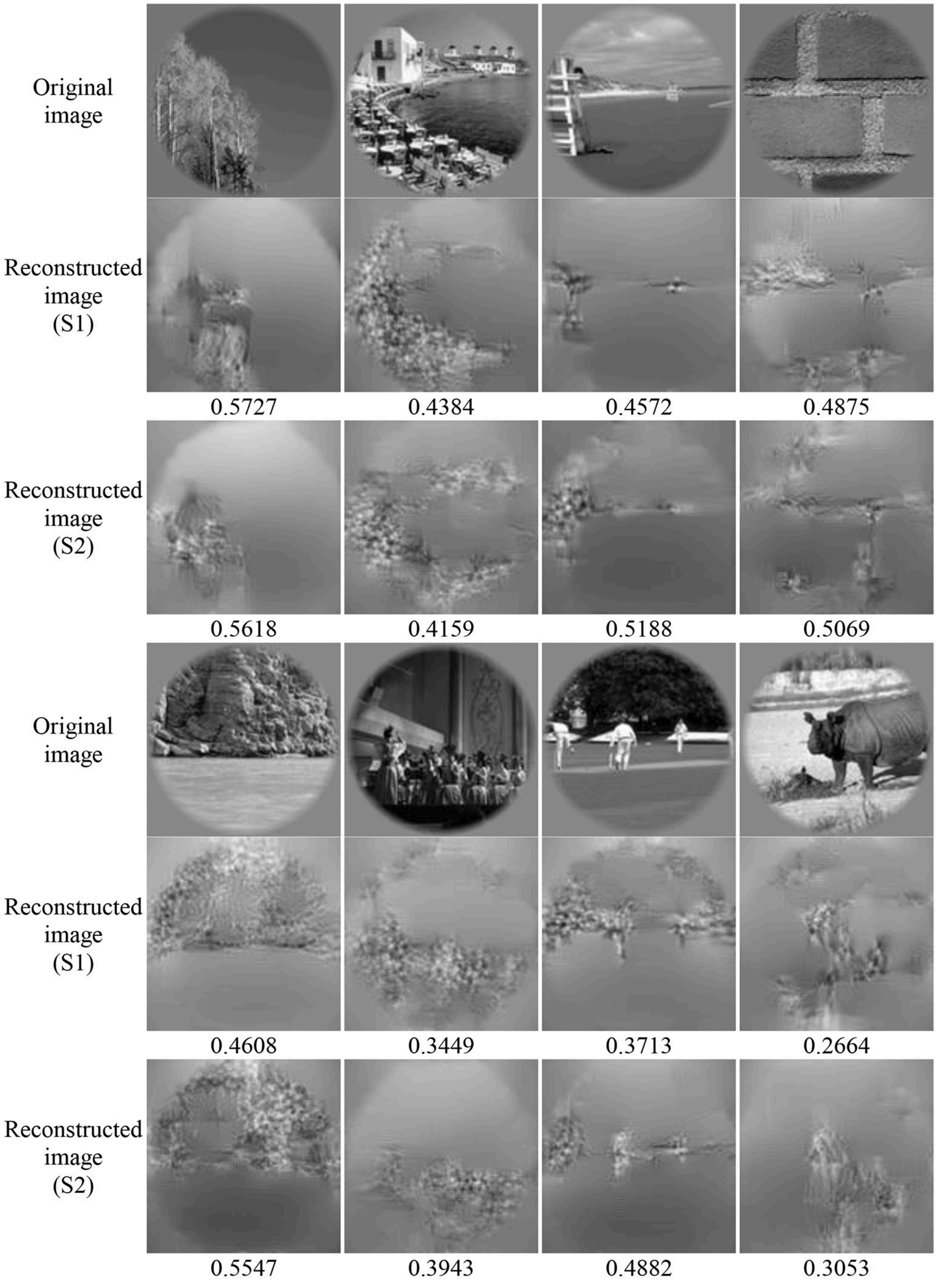

FIGURE 5 | Performance of image reconstruction. The numbers below the reconstructed image represent the accuracy of the reconstructions assessed by CWSSIM.

set. Each reconstruction was assessed by each rater once and 3120 comparisons in total were presented.

RESULTS

Comparative Analysis of ROMP and YALL1 for Decoding Model

ROMP and YALL1 were used to solve the decoded model, and the prediction accuracy of several typical layer features decoded from the fMRI data of S1 is shown in **Figure 2**. The average prediction accuracy of ROMP in pool1, pool2, conv3, conv4, and conv5 reached 0.266, 0.183, 0.283, 0.263, and 0.239 respectively, which were significantly higher than that of YALL1 on t-statistics at a significance level of 10^{-5}. This finding might indicate that ROMP better reflected the sparsity of visual perception. Furthermore, ROMP was much faster than YALL1, which was particularly important for tens of thousands of CNN features. In conclusion, we finally selected ROMP as our decoding model solution.

Prediction Accuracy of Different Layers of CNN

We calculated the prediction accuracy of all layers of CNN based on both the fRMI data of S1 and S2. As shown in **Figure 3**, higher prediction accuracy was obtained in pool1, conv2, conv3, conv4, conv5, fc6, fc7, and fc8 layers, whereas the prediction accuracy in conv1, relu1, relu2, relu3, relu4, relu5, relu6, and relu7 layers was low possibly because the ReLU function reduced the predictability of the linear decoding model. Intuitively, image reconstruction performed better when it utilized the features of the layer with higher prediction accuracy as image representations. However, under the same prediction accuracy, the reconstruction method used in this paper had better accuracy rate in the lower layer because the more distortion was generated during back propagating the higher layer features. Thus, we finally selected pool1 layer as the image representation Φ (see section Reconstructing Image From CNN Features for more details) to reconstruct the original image from the voxel response.

Contribution of Each Visual Area When Decoding fMRI Signals Into Each Layer Feature

The visual area assignments of the significant voxels across all CNN layers are shown in **Figure 4**. The results showed that the assignments of the significant voxels in V1 and V2 had a decreasing trend (Mann–Kendall test, $p < 0.05$) with the CNN layer, whereas the assignments of the significant voxels in V3 and V4 had an increasing trend (Mann–Kendall test, $p < 0.05$) for both subjects. That is, most significant voxels assigned to shallow convolutional layers were located in early visual areas, whereas most significant voxels assigned to deep convolutional layers were located in downstream visual areas. As we know, CNN is hierarchically organized with feature complexity. Thus, these findings provided quantitative evidence again for the thesis that the visual ventral stream was hierarchically organized (Markov et al., 2014), with downstream areas processing increasingly complex features of the retinal input.

Performance of Image Reconstruction

Image reconstruction was implemented on the validation set based on the pool1 features decoded from the voxel responses using the decoding model trained in the training set. Part of the original images of the validation set and the corresponding reconstructed images are shown in **Figure 5**. Most reconstructed images were found to clearly capture the position, shape, and even the texture information of the object in the original image in the case that the stimuli were grayscale. Moreover, we found that most of reconstructed images reproduced foreground objects well but were less sensitive to perceptually less salient objects or backgrounds. To some extent, this finding showed that the visual perception of the brain measured by fMRI was selective during image understanding, which might be the main reason why reconstruction images tended to regenerate those image parts relevant to visual perception. In addition, accuracy of the reconstructions was assessed by CWSSIM and human judgment (**Figure 6**). The average accuracy of the reconstructions for S1 and S2 reached 0.3921 (80.1% by human judgment) and 0.3938 (70.4% by human judgment), respectively. In addition, both S1 and S2 were significantly more accurate than chance ($p < 10^{-5}$, T-test). As a way that computers can judge, the accuracy of the

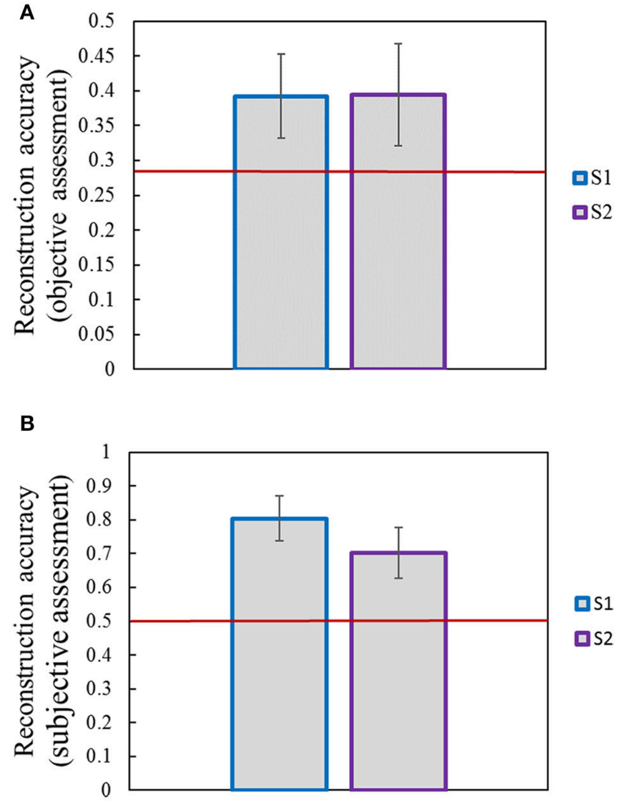

FIGURE 6 | Accuracy of reconstructions. **(A)** Accuracy of reconstructions assessed by objective method (CWSSIM). **(B)** Accuracy of reconstructions assessed by subjective method. Raters were presented with an original image and had to choose its reconstruction out of two. The red line indicates chance performance. Error bars show the standard deviation of the mean accuracy.

reconstructions assessed by CWSSIM may be not consistent with the judgment of humans.

DISCUSSION

Same as most studies, the linear model was selected in this paper to decode fMRI signals to CNN features. A more complex model inspired by visual mechanism may be able to improve the decoding effect, including the Gabor wavelet pyramid model (Kay et al., 2008) to predict the responses of voxels in early visual areas. We compared the ROMP and YALL1 to select a better one to solve the decoding model and found that sparsity was a good feature for solving algorithms. However, we just compared two typical algorithms in two classes of sparse optimization methods. Thus, better algorithms need to be explored to improve the decoding accuracy.

The existing studies mostly only analyzed the relationship between the response of visual voxels and the hierarchical features of the five convolutional layers of CNN and three fully connected layers. In this paper, the decoding accuracy of all layers in CNN was analyzed (**Figure 3**) and some interesting phenomena that have not been discovered before were found. For example, the prediction accuracy of convolutional and fully connected layers were relatively higher (except for conv1), which might be the reason why most studies only analyzed them. We found that the prediction accuracy of conv1 layer was low but that of the next layer pool1 was higher. The reason might be that the fMRI signals were more like pool1, which reflected the responses of a group of nerve cells rather than a single nerve cell, thus it was constrained to decode fMRI signals into the lowest level but most sophisticated features in the first layer of CNN. Moreover, all the layer features of ReLU had relatively low prediction accuracy. As we know, ReLU function is originally an approximate simulation of the activation model of brain neurons for faster and better training of a deeper network model. This phenomenon may lead to the characteristics of the relu layer deviating from the visual perception process of the human brain (measured by fMRI). These findings based on fMRI may be useful for the improvement of CNN.

Recently, several research found the similarity between CNN and the visual pathway through visual encoding (Agrawal et al., 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015a,b; Khaligh-Razavi et al., 2016) or decoding (Horikawa and Kamitani, 2017a,b; Wen et al., 2017). These findings were verified more carefully in this paper through the analysis of the contribution of each visual area during the decoding of fMRI signals to all the layer features of CNN (**Figure 4**). From another point of view, these cases may be because CNN is closer to the human brain in image understanding, thus it can achieve various essential improvements in image target recognition and detection and other functions.

In the final process of image reconstruction, we obtained better reconstruction performance by inverting pool1 layer feature decoded from fMRI signals, although the prediction accuracy of pool1 was not the highest. We tried to reconstruct images based on the high-level layer with higher accuracy (such as that of conv3) but did not work well probably due to the ultimate goal of the CNN to identify the target in images. Thus, the higher layer features contained more semantic information and less low-level features of images. This case led to larger distortion in the reconstructed images by inverting higher layer features even when the features were extracted directly from the original image (Mahendran and Vedaldi, 2015, 2016). In the case of similar prediction accuracy, better reconstruction could be implemented based on the pool1 layer but could also lead to the recovered information that are almost low-level information, such as location, edge, texture, and so on. To achieve better image reconstruction performance, a fusion method based on CNN multi-layer features rather than single-layer features is encouraged. In this way, the details of the image can be recovered better by using the low-level layers of CNN, whereas the semantics of the image can be guaranteed by the high-level features.

CONCLUSION

This paper presents a novel method for reconstructing constraint-free natural images from fMRI signals based on CNN. Different from direct reconstruction from fMRI signals, we transferred the understanding of brain activity into the understanding of feature representation in CNN by training a mapping from fMRI signals to hierarchical features extracted from CNN. Thus, image reconstruction from fMRI signals became the problem of CNN feature visualization. By iteratively optimizing to find the matched image, we finally achieved the promising results for the challenging constraint-free natural image reconstruction. Furthermore, the homology of human and machine visions was validated based on the experimental results. As the semantic prior information of the stimuli were not used when training decoding model, any category of images (not constraint by the training set) could be reconstructed theoretically based on the CNN pre-trained on the massive samples of ImageNet. To achieve better image reconstruction performance on colorful images or videos, CNN multi-layer features representing different levels of image features should be taken into account.

AUTHOR CONTRIBUTIONS

This paper was accomplished by all the authors with the assignments that CZ, LW, and BY conceived the idea; CZ and KQ performed the analysis; CZ, KQ, LW, LT, and YZ co-wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under grant 2017YFB1002502 and the National Natural Science Foundation of China (No. 61701089, No. 61601518 and No. 61372172).

REFERENCES

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J. L. (2014). Pixels to voxels: modeling visual representation in the human brain. arXiv:1407.5104.

Blunsom, P., Grefenstette, E., and Kalchbrenner, N. (2014). "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, MD).

Brooks, A. C., Zhao, X., and Pappas, T. N. (2008). Structural similarity quality metrics in a coding context: exploring the space of realistic distortions. *IEEE Trans. Image Process.* 17, 1261–1273. doi: 10.1109/TIP.2008.926161

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6:27755. doi: 10.1038/srep27755

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/S1053-8119(03)00049-1

Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., et al. (2013). "Recent advances in deep learning for speech research at Microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*: *IEEE* (Vancouver, BC).

Denton, E., Chintala, S., Szlam, A., and Fergus, R. (2015). Deep generative image models using a laplacian pyramid of adversarial networks. *Comp. Sci.*

Donoho, D. L., and Stark, P. B. (1989). Uncertainty principles and signal recovery. *SIAM J. Appl. Math.* 49, 906–931. doi: 10.1137/0149053

Du, C., Du, C., and He, H. (2017). Sharing deep generative representation for perceived image reconstruction from human brain activity. arXiv:1704.07575.

Güçlü, U., and van Gerven, M. A. (2015a). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Güçlü, U., and van Gerven, M. A. (2015b). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *Neuroimage.* 145(Pt B), 329–336. doi: 10.1016/j.neuroimage.2015.12.036

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

Horikawa, T., and Kamitani, Y. (2017b). Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Front. Comput. Neurosci.* 11:4. doi: 10.3389/fncom.2017.00004

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science* 340, 639–642. doi: 10.1126/science.1234330

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, FL: ACM).

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature* 452, 352–355. doi: 10.1038/nature06713

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., and Kriegeskorte, N. (2016). Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* 76, 184–197. doi: 10.1016/j.jmp.2016.10.007

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," *in Advances in Neural Information Processing Systems* (Lake Tahoe).

Le, Q., and Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:*1405.4053.

Mahendran, A., and Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vis.* 120, 233–255. doi: 10.1007/s11263-016-0911-8

Mahendran, A., and Vedaldi, A. (2015). "Understanding deep image representations by inverting them," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA).

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., et al. (2014). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *J. Comp. Neurol.* 522, 225–259. doi: 10.1002/cne.23458

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science* 320, 1191–1195. doi: 10.1126/science.1152876

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M. A., Morito, Y., Tanabe, H. C., et al. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60, 915–929. doi: 10.1016/j.neuron.2008.11.004

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron* 63, 902–915. doi: 10.1016/j.neuron.2009.09.006

Needell, D., and Vershynin, R. (2009). Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations Comput. Math.* 9, 317–334. doi: 10.1007/s10208-008-9031-3

Needell, D., and Vershynin, R. (2010). Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE J. Sel. Top. Signal Process.* 4, 310–316. doi: 10.1109/JSTSP.2010.2042412

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* 21, 1641–1646. doi: 10.1016/j.cub.2011.08.031

Postle, B. R. (2015). The cognitive neuroscience of visual short-term memory. *Curr. Opin. Behav. Sci.* 1, 40–46. doi: 10.1016/j.cobeha.2014.08.004

Reddy, L., Tsuchiya, N., and Serre, T. (2010). Reading the mind's eye: decoding category information during mental imagery. *Neuroimage* 50, 818–825. doi: 10.1016/j.neuroimage.2009.11.084

Schoenmakers, S., Barth, M., Heskes, T., and Van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *Neuroimage* 83, 951–961. doi: 10.1016/j.neuroimage.2013.07.043

Schoenmakers, S., Güçlü, U., Van Gerven, M., and Heskes, T. (2014a). Gaussian mixture models and semantic gating improve reconstructions from human brain activity. *Front. Comput. Neurosci.* 8:173. doi: 10.3389/fncom.2014.00173

Schoenmakers, S., Van Gerven, M., and Heskes, T. (2014b). "Gaussian mixture models improve fMRI-based image reconstruction," in *Pattern Recognition in Neuroimaging, 2014 International Workshop on IEEE* (Tübingen).

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., et al. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage* 33, 1104–1116. doi: 10.1016/j.neuroimage.2006.06.062

Vedaldi, A., and Lenc, K. (2015). "MatConvNet: convolutional neural networks for MATLAB," in *Proceedings of the 23rd ACM International Conference on Multimedia* (Brisbane, QLD: ACM).

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* 1–25. doi: 10.1093/cercor/bhx268

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

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yang, J., and Zhang, Y. (2011). Alternating direction algorithms for L1-problems in compressive sensing. *SIAM J. Sci. Comput.* 33, 250–278. doi: 10.1137/090777761

Yargholi, E., and Hossein-Zadeh, G. A. (2015). Reconstruction of digit images from human brain fMRI activity through connectivity informed Bayesian networks. *J. Neurosci. Methods*. 257, 159–167. doi: 10.1016/j.jneumeth.2015.09.032

Yargholi, E., and Hossein-Zadeh, G. A. (2016). Brain decoding-classification of hand written digits from fMRI data employing bayesian networks. *Front. Hum. Neurosci.* 10:351. doi: 10.3389/fnhum.2016.00351

Zafar, R., Malik, A. S., Kamel, N., Dass, S. C., Abdullah, J. M., Reza, F., et al. (2015). Decoding of visual information from human brain activity: A review of fMRI and EEG studies. *J. Integr. Neurosci.* 14, 155–168. doi: 10.1142/S0219635215500089

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.