Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

Functional diversity of visual cortex improves constraint-free natural image reconstruction from human brain activity

Lingxiao Yang[a], Hui Zhen[b], Le Li[c], Yuanning Li[d], Han Zhang[d], Xiaohua Xie[a], Ru-Yuan Zhang[e,f,g,*]

[a] School of Computer Science and Engineering, Sun Yat-sen University, Guangdong 510006, China
[b] Bytedance Inc, Beijing 100190, China
[c] School of Computer Science, University of Waterloo, Ontario N2L3G1, Canada
[d] School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China
[e] Institute of Psychology and Behavioral Science, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 20030, China
[f] Shanghai Mental Health Center, School of Medicine, Shanghai Jiao Tong University, Shanghai 20030, China
[g] Key Laboratory of Brain-Machine Intelligence for Information Behavior- Ministry of Education, Shanghai International Studies University, Shanghai 200083, China

ARTICLE INFO

Article history:
Received 6 March 2023
Received in revised form 29 July 2023
Accepted 31 August 2023
Available online xxx

Keywords:
fMRI decoding
Brain-computer interfaces
Brain function-related model
Conditional generative network
Natural image reconstruction

ABSTRACT

Previous brain decoding studies using functional magnetic resonance imaging (fMRI) have greatly advanced our understanding of human visual coding and non-invasive brain-machine interfaces. However, most of these studies focus on classifying a limited number of image categories or reconstructing visual images with additional information, *e.g.*, semantic categories and textual cues. Constraint-free visual reconstruction remains scarce. Here, we propose a generative network based on the functional diversity of the human visual cortex (FDGen) that takes multivariate brain activity as input and directly reconstructs natural images perceived by observers without any additional cues (semantic categories or textual description). Our FDGen is augmented by two bio-inspired computational modules. Based on the functional specializations of the human visual cortex, we propose a new function-based input module (FIM) that projects responses from different brain regions into separate feature spaces. Second, inspired by human attention, we construct a computational module to derive attentive feature weights at the function level to refine the feature map. These function-selection modules (FSMs) allow the network to dynamically select multiscale visual information during the generation process. We test FDGen on the popular fMRI datasets of natural images and achieve highly robust performance. Our work represents an important step forward in the development of fMRI-based brain decoding algorithms and highlights the utility of neuroscience theories in the design of deep learning models.

1. Introduction

The movie "*The Matrix*" depicts a magical brain-machine interface that can accurately decode and willfully control human mental states. Although fictional, high-fidelity brain-machine interfaces have long fascinated neuroscientists. Brain decoding, as one of the most important techniques in brain-machine interfaces, not only advances our theoretical understanding of cortical processing, but also has tremendous transnational value in medicine [1]. Electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI) are two of the most widely used non-invasive neuroimaging techniques [2–5]. Since our task in this paper is the pixel-level reconstruction of static images, we focus on fMRI-based brain activity because fMRI provides greater spatial detail of visual processing in the human brain than EEG [6].

However, the majority of early visual decoding models focus on pattern classification - these models take multidimensional brain activity as input and output labels of either simple visual features or image categories [7–10]. Recently, several studies on image reconstruction from brain activity have been published. Some of these works focus on reconstructing either simple visual patterns, such as checkerboard patterns [11], or realistic images in specific domains, such as handwritten digits [12] or faces [13]. Images from these domains often contain narrowly defined statistical properties that can be used as priors to achieve relatively good results [14]. Some recent studies bypass image reconstruction methods and instead seek to use generative similar images with additional texture cues [15].

Although several studies have been proposed to reconstruct perceived natural images based on fMRI signals [16–21], constraint-free image reconstruction is still challenging because it requires accurate retrieval of multi-scale visual information. In addition, full image reconstruction is remarkably valuable in the fields of brain diagnosis and computer vision. First, brain decoding has been widely used to address

[*] Corresponding author.
E-mail address: ruyuanzhang@sjtu.edu.cn (R.-Y. Zhang).

https://doi.org/10.1016/j.fmre.2023.08.010

various theoretical problems in medical diagnosis. For example, accurate brain decoding enables the identification of preserved brain networks in patients with disorders of consciousness, thus promoting more accurate diagnosis and prognosis [22]. Second, advances in brain decoding also provide insights into computational models for other applications, such as computer vision and image processing [23–25].

Accurate image reconstruction from brain activity faces three major challenges. First, fMRI signals are notoriously noisy, encompassing the effects of various non-neural factors, such as thermal noise, physiological and vascular noise, and head motion [26]. The noise and other uncontrolled factors fundamentally limit the quality of fMRI data and set the upper performance limit that the best decoding model can achieve. How to maximally attenuate these non-neural artifacts is still an active research field in neuroimaging. Second, due to the considerable cost of fMRI experiments, fMRI-based image reconstruction is hampered by the practical barrier of obtaining large amounts of training data. The complex statistical regularities of natural images may require sophisticated nonlinear models, which in turn require large amounts of images and corresponding brain data to learn their statistical dependence. Third, even if the quality and quantity of fMRI data is guaranteed, the development of accurate decoding algorithms may still be difficult because it relies heavily on our knowledge of the encoding mechanisms in the brain [27]. Unfortunately, such quantitative models are not yet well established.

In this paper, we focus on the third challenge of developing a novel generator based on the functional diversity of brain regions (FDGen)to improve the performance of brain reading. Our FDGen is inspired by some well-known neuroscience theories and is trained together with a discriminator to directly map multivariate voxel responses to corresponding natural images. Specifically, FDGen contains two components: a function-based input module and a function-selection module. The function-based input module is designed using functional parcellations of the human visual cortex [28]. It decomposes the whole brain activation vector into different parcels of region-of-interests according to their brain functions, and transforms them from the raw activity space into different feature spaces. Compared to the conventional vector-based input module, our function-based input module explicitly takes into account the different feature processing mechanisms of brain regions. Our function-selection module inherits the attentional mechanisms in machine learning. This module learns the importance of different brain functions to selectively process the most informative signals during the reconstruction process. Experimental results show that the proposed FDGen significantly improves the reconstruction performance in both qualitative and quantitative comparisons, demonstrating the usefulness of brain function partitioning in image reconstruction. More importantly, the training and test images belong to different categories in our experiments. It clearly confirms that FDGen has a good generalization ability for natural image reconstruction.

2. Methods and materials

Our framework follows the general pipeline of conditional GANs [29–31]. The main contribution of this paper is a redesigned generator (Fig. 1) inspired by the functional specializations of the human visual cortex. In a typical image reconstruction problem, the goal of this paper is to learn a mapping network to reconstruct an input image stimulus $\boldsymbol{X}_i$ (*i.e.*, an RGB image) from the high-dimensional brain activity vector $\boldsymbol{v}_i$ (*i.e.*, responses of many voxels).

2.1. Our functional diversity based generator

Function-based input module (FIM). Considering the different behavior of human brain areas, we take into account different functional brain areas and propose a new input module called Function-based Input Module. Our FIM uses additional region-of-interest masks, which can be easily obtained via retinotopic or functional localization experiments [32], and are readily available in almost all natural image fMRI datasets. In FIM, we first use a function-split layer to partition each high-dimensional brain activity vi into a few function-wise activities $\boldsymbol{v}_i^r,\ r = (1, 2, \dots, \boldsymbol{R})$ with respect to the brain region-of-interest masks provided in a dataset. i is the data instance index. FIM then transforms each parceled activity into a separate space with R different function f^r and then aggregates all embedded features via a fully connected layer. The output is followed by a reshape operator to generate the initial 3-dimensional feature plane. Fig. 1d shows the process, which can be formulated as follows:

$$\boldsymbol{v}_i^r = \boldsymbol{v}_i \odot \boldsymbol{m}^r \tag{1}$$

$$\boldsymbol{y}_i^r = f^r\left(\boldsymbol{v}_i^r\right) \tag{2}$$

$$\boldsymbol{Y}_i = f\left(\boldsymbol{y}_i^r\right) \tag{3}$$

Here, $\odot$ is the element-wise product to filter r functional responses by the supplied brain region mask $\boldsymbol{m}^r$. Each mask contains binary values and has the same dimension as the brain activity $\boldsymbol{v}_i$. f^r is a function-wise mapping to transform each functional response into a feature space. We instantiate f^r as two fully connected layers for each input activity. To simplify the process, we fix all output dimensions of $\boldsymbol{y}_i^r \in \mathcal{R}^{[4096/\boldsymbol{R}]}$, where [4096/R] = ceil(4096/7) = 586 on the Shen's dataset. In addition, f is another fully connected layer, followed by a reshape layer to form the initial feature plane $\boldsymbol{Y}_i \ \in \mathcal{R}^{256\times4\times4}$. 4×4 is the feature spatial size, and 256 is the feature channels. This implementation, which employs three fully connected layers to map brain activity to the initial 3D features, is consistent with conventional vector-based input module as shown in Fig. 1c [16], but has the following advantages: the activity in each brain region is selectively processed by f^r corresponding to that function, and f^r can be instantiated differently for different functions. As a result, our FIM allows the processing of features from different brain functions via different f^r.

Function-selection module (FSM). We propose a function-selection module (FSM) to refine the intermediate layers, as shown in Fig. 1a. The logic is that, to reconstruct certain types of visual information (*e.g.*, texture), it is better to emphasize the information in the activity of the corresponding brain functions (*e.g.*, V2 and V3), and downplay the activity in other regions. Such a selective reweighting of functional responses may allow to extract the most useful visual information and improve the reconstruction performance. Specifically, in our FSM block, we first upsample the input features with a deconvolutional layer and then refine the upsampled features with a convolutional layer. As all generator blocks have similar operators, we omit the block index in the following for simplicity and define all operations for a single block as an example. For each block, the upsampling process can be formulated as:

$$\tilde{\boldsymbol{Y}}_i = Conv\left(DeConv\left(\boldsymbol{Y}_i\right)\right) \tag{4}$$

where $\boldsymbol{Y}_i$ is current block input. We then generate per-function weights $\boldsymbol{w}_i \in \boldsymbol{R}^R$ from the current feature planes $\tilde{\boldsymbol{Y}}_i$. This is achieved by a spatial global average pooling (GAP) layer and two fully connected layers to capture attentional cues based on the current features. Formally, $\boldsymbol{w}_i$ is obtained by:

$$\boldsymbol{w}_i = softmax\left(\boldsymbol{W}_2 * \delta\left(\boldsymbol{W}_1 * \frac{1}{W\times H}\sum_{w=1}^{W}\sum_{h=1}^{H}\tilde{\boldsymbol{Y}}_i\right)\right), \tag{5}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are trainable parameters in the two fully connected layers respectively. $*$ is the matrix or vector product multiplication. δ is a leaky ReLU with slope 0.3. After obtaining the per-function weights, we then apply them to the brain function-wise features $\boldsymbol{y}_i^r$ (Eq. 2) (see dash arrow in Fig. 1) to emphasize important features. Then the new features are obtained by aggregating all the weighted function-wise features. The above processes can be formulated as:

$$\hat{\boldsymbol{y}}_i = \sum_{r=1}^{R} w_i^r \boldsymbol{y}_i^r \tag{6}$$

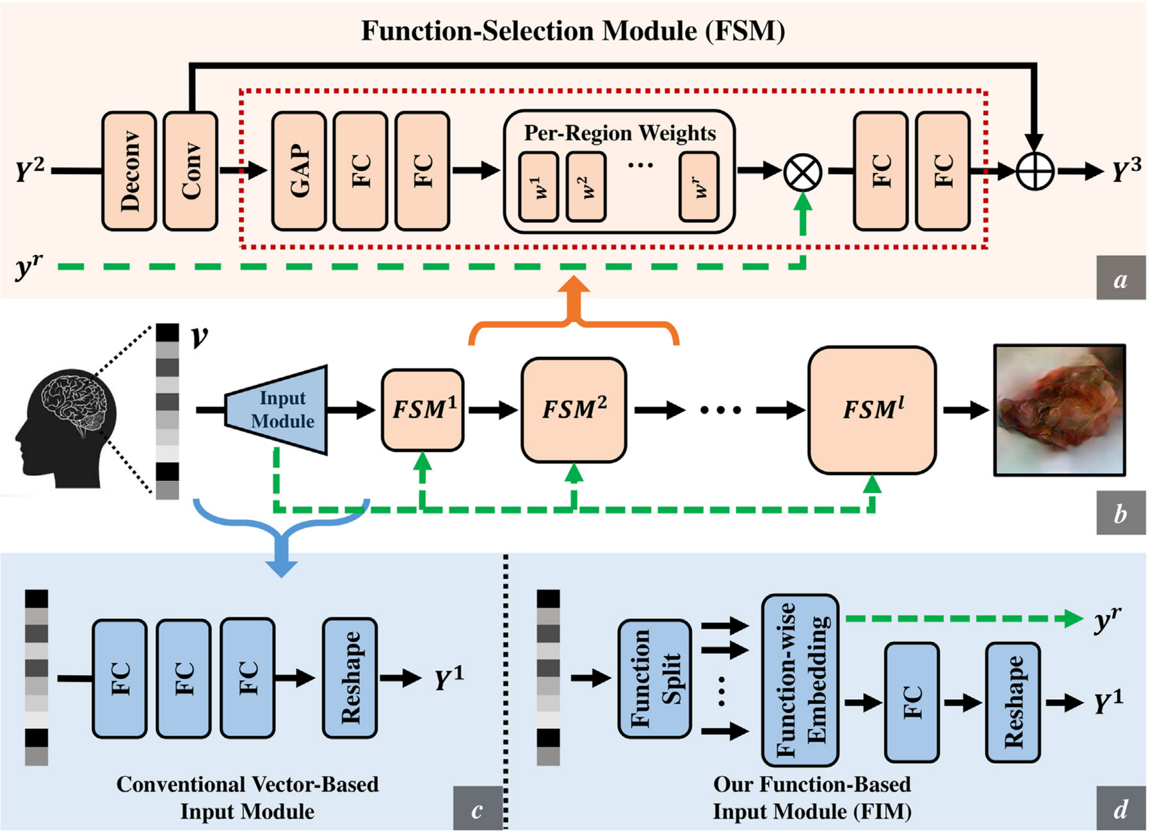

Fig. 1. An overview of our generator (FDGen) is shown in b. FDGen consists of a function-based input module (FIM, d) and a multi-stacked function-selection module FSM (see the full part in a) to transform the input brain activity v into natural images. Unlike the conventional vector-based input module (c) used in other methods, in FIM, we introduce a function-split layer and a function-wise embedding layer to first extract the functional information from brain responses. Then a fully connected layer and a reshaping layer are attached to obtain the initial feature plane. Our FSMs automatically learn function weights, and these weights are incorporated into current feature planes to refine the generation process.

Here, $\hat{\boldsymbol{y}}_i \in \boldsymbol{R}^{586}$ are the aggregated features. Finally, we combine this new type of feature with $\tilde{\boldsymbol{Y}}_i$ through a residual connection to produce current block's output $\widehat{\boldsymbol{Y}}_i$:

$$\widehat{\boldsymbol{Y}}_i = Expand\left(\boldsymbol{W}_4 * \delta\left(\boldsymbol{W}_3 * \hat{\boldsymbol{y}}_i\right),\ \tilde{\boldsymbol{Y}}_i\right) + \tilde{\boldsymbol{Y}}_i \tag{7}$$

Here, $Expand(\mathrm{x}, \mathrm{z})$ means expanding the input x to have the same spatial size as z. Note that, Eq. 7 uses the attended features $\hat{\boldsymbol{y}}_i$ to refine the $\tilde{\boldsymbol{Y}}_i$, which has a large impact on the final reconstruction performance as our experiment showed. Currently, our FSM only uses different brain regions to refine feature channels, it may be more suitable to simultaneously refine the spatial and feature channels, which could be an interesting research direction in the future.

2.2. Model architecture

Generator. Our generator is based on the above two modules and can be formulated as: FIM-FSM256-FSM512-FSM128-FSM64-FSM32-FSM3. The FIM is: FC[4096/R]R-FC[4096/R]R-Concat-FC4096-Reshape. The FCR is R parallel fully connected layers connected to R brain regions (*e.g.*, V1, V2). The output of each fully connected layer has channels of [4096/R], where [·] is the ceiling operator. Our FSM is formulated in Eqs. 4–8. In particular, both $\boldsymbol{W}_1$ and $\boldsymbol{W}_3$ in Eqs. (6) and (7) produce the same number of input channels. $\boldsymbol{W}_2$ produces R functional weights while $\boldsymbol{W}_4$ adapts the feature channels to be consistent with $\tilde{\boldsymbol{Y}}_i$. The final FSM block outputs a 3-channel image for loss computation during training. All deconvolutional and convolutional layers have a nonlinear leaky ReLU function with a slope of 0.3.

Discriminator. Our goal in this paper is to design an advanced generator to improve the reconstruction performance. Here, we use a simple convolutional network [16] as our discriminator. This discriminator D consists of: C32-C64-C128-C256-C256-AP-DropOut0.5-FC256-Dropout0.5-FC2. "CX" and "FCX" are the convolutional layers and the fully connected layers with X output channels respectively. AP is the average pooling layer. DropOut is used to reduce the risk of overfitting. In addition, the nonlinear function - ReLU is applied after each convolutional layer and after the first fully connected layer. The kernel sizes and strides of the convolutional layers are set to [7, 5, 3, 3, 3] and [4, 1, 2, 1, 2], respectively, from the first to the fifth layer. The output of this network is fed into a softmax to decide whether the input image is real or fake.

Loss function. Given a training database of many pairs of samples - $(\boldsymbol{v}_i, \boldsymbol{X}_i),\ i = (0, 1, \ldots, n)$ and brain region-of-interest masks $\boldsymbol{m}$, our goal is to train the generator G to fool the discriminator D with the adversarial loss L_{adv}, as well as two additional constraints (the image reconstruction loss L_{img} and the feature comparison loss L_{feat} [16]. The overall objective function for training G is a linear combination of the three losses above:

$$L_G = \sum_i^N \left\{ \lambda_{img} \underbrace{||\boldsymbol{X}_i - G(\boldsymbol{v}_i, \boldsymbol{m})||_2^2}_{L_{img}} + \lambda_{feat} \underbrace{||C(\boldsymbol{X}_i) - C(G(\boldsymbol{v}_i, \boldsymbol{m}))||_2^2}_{L_{feat}} - \lambda_{adv} \underbrace{\log\left(D\left(G\left(\boldsymbol{v}_i, \boldsymbol{m}\right)\right)\right)} \right\} \tag{8}$$

Both L_{img} and L_{feat} are L_2 regression losses that force the generated image G($\boldsymbol{v}_i$, $\boldsymbol{m}$) to approximate the original image $\boldsymbol{X}_i$ in pixel and feature space, respectively. For the feature comparison loss L_{feat}, we use the

Fig. 2. Eight example pictures from the Shen's dataset [16]. This dataset contains many categories with different backgrounds.

last pooling layer (pool5) of AlexNet [33] to compute feature distances. Similar to Shen, et al. [16], the AlexNet here is pre-trained on ImageNet [34] and is not tuned throughout the training process.

The output of discriminator D is fed into a 2-way softmax layer to discriminate the original stimulus image from the reconstructed image with the following formulation:

$$L_{disc} = -\sum_{i}^{N} \log\left(\mathrm{D}\left(\boldsymbol{X}_i\right)\right) + \log\left(1 - D\left(G\left(\boldsymbol{v}_i, \boldsymbol{m}\right)\right)\right) \quad (9)$$

During training, G and D are optimized iteratively to compete with each other via Eq. (8). After training, the trained generator G is used to reconstruct an image with the input brain activity vi and the region-of-interest masks $\boldsymbol{m}$.

Training settings. Our FDGen is trained separately for each subject because the dimensions of the input brain activity vary across subjects, which is a common setting in previous methods. Specifically, during training, the input images are first resized to 256 × 256 and then randomly cropped to 224 × 224. All variants of our models are optimized by the Adam solver [35] with a batch size of 64 on 4 GPUs, momentum $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Additionally, the batch mean and variance for the batch normalization layer (after each convolutional layer) are computed within each GPU. The learning rate is set to 0.0002 for the first 150 epochs. We then reduce the learning rate to 0.00002 for the last 50 epochs. During testing, we average 24 trials to reduce noise [16]. λ_{img}, λ_{feat}, and λ_{adv} in Eq. (9) are set to 100, 25, and 1, respectively.

Evaluation methods. Following previous studies [17,36], we report Pearson correlation coefficient (PCC) and structural similarity index (SSIM) values for pairwise similarity comparisons. PCC measures the correlation between a generated image and a target image at the pixel level. For PCC, we first transform both generated and target images into 1-dimensional vectors (for multi-channel images, pixels of different color channels are concatenated into one vector), and then compute the PCC between the two vectors. SSIM measures the feature-level similarity between two images. For multi-channel images, we compute the SSIM score in each channel independently and average the scores across channels to obtain the final SSIM score. Each reconstructed image is compared with two candidate images: one is always the groundtruth image, and the other is randomly selected from the rest of the test images. A binary decision is made to select the candidate image that has a higher similarity score in both evaluation metrics (*i.e.*, PCC or SSIM). This process is performed for all possible candidate images other than the groundtruth image in the test dataset. The final reconstruction accuracy of this generated image is indexed by the percentage of comparisons in which the groundtruth image wins. This evaluation examines the extent to which a reconstructed image is more similar to the groundtruth image than others. We call it PCC-c or SSIM-c for simplicity. In addition to the above comparisons, we also include self-comparison metrics that are widely used in computer vision. Specifically, we include the LPIPS evaluation metric [37]. LPIPS uses a deep model pre-trained on a large-scale dataset labeled based on human perception to evaluate the similarity between a synthetic image and a real image. The results in Zhang, et al. [37] show that LPIPS is more consistent with human judgment than some low-level perceptual metrics (*e.g.*, SSIM). Note that higher PCC and SSIM scores indicate better performance, while lower LPIPS scores indicate better performance. In the following sections, we refer to the self-comparison metrics as PCC-s, SSIM-s, and LPIPS-s.

Dataset. We conduct experiments on the popular publicly available benchmark [16,17] (referred to as Shen's dataset). This dataset provides image stimuli and corresponding fMRI data. The full set of images in Shen's dataset contains four parts: artificial shapes, alphabetic letters, training, and test natural images. We used only the natural images and corresponding fMRI data in our experiments. The natural image data includes 1250 natural images drawn from 200 selected categories in ImageNet. 1200 images from 150 categories are used as the training data, and the remaining 50 images from other 50 categories are used for testing. We emphasize that the image categories in the training and test sets do not overlap. Such a non-overlapping design imposes additional difficulties on image reconstruction because it excludes the effects of semantic priors for reconstruction. This non-overlapping design also explores the cross-category generalization of reconstruction models. Shen's dataset includes fMRI data from three subjects. For each subject, each image in the training and the test sets is presented for 5 trials and 24 trials respectively. Therefore, the total number of data pairs for each subject in the training and test sets is 6250 and 1200, respectively. In addition, this dataset provides brain function masks (V1/V2/V3/V4/LOC/FFA/PPA). All of these masks are used in our experiments. Fig. 2 shows some images from this dataset.

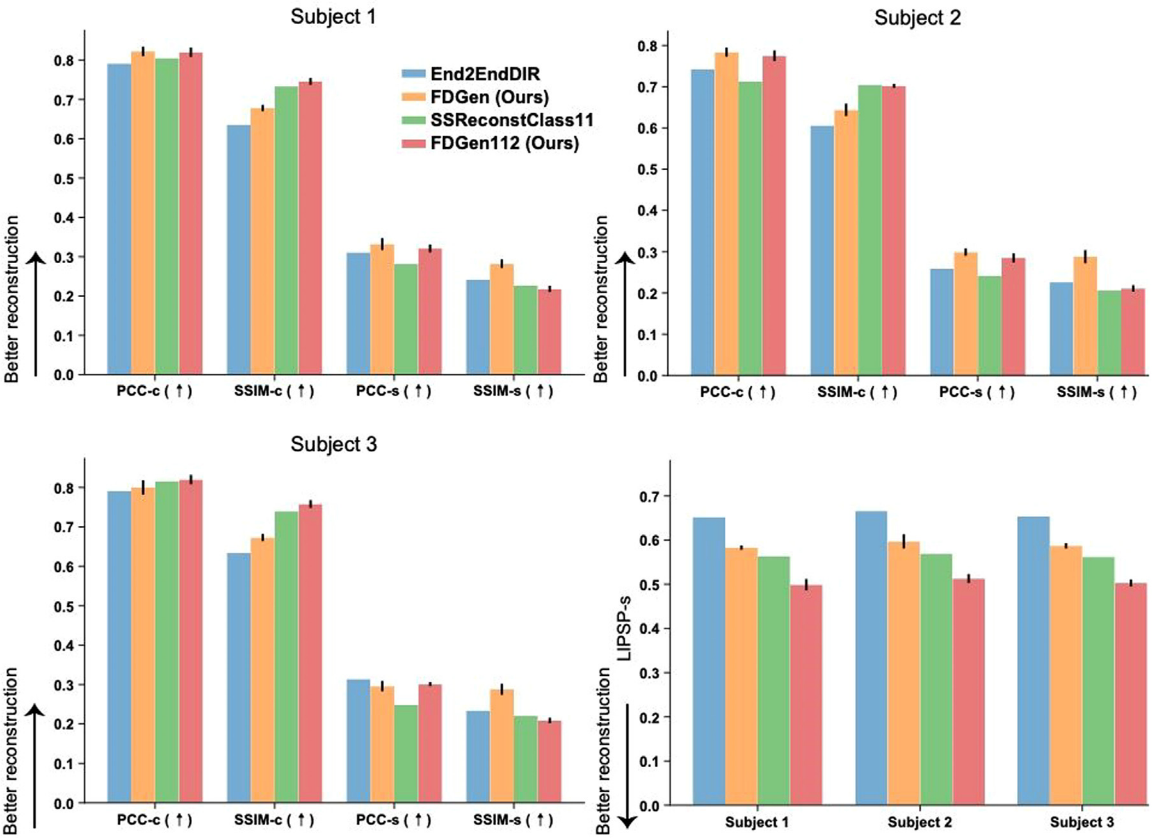

Fig. 3. Reconstruction performance of different methods on the Shen's dataset. Details of each evaluation metric are presented in the main text. For these metrics, ↑ indicates a higher value is better, while means a lower is better. "112" means using the image size of 112 for reconstruction.

3. Results

3.1. *FDGen exhibits robust brain decoding performance as compared to previous models*

We compare our FDGen with several previous reconstruction methods. There are two popular fMRI datasets on the generic object decoding task (different fMRI recordings with the same natural images), Shen's dataset [16] and Kamitani's dataset [38]. The main difference between these datasets is the number of fMRI recordings for each training image. Kamitani's dataset has only 1 trial per image, whereas Shen's dataset has 5 trials per image. As a result, Shen's dataset can reduce the fMRI noise through trial averaging and is used to evaluate our method. In addition, we found that many previous methods performed model comparisons using different fMRI datasets. For example, End2EndDIR [16] and Ren, et al. [39] evaluated their methods on the Shen's datasets, while ssfmri2im [19] and its journal version - SSReconstClass [24], and Fang, et al. [20] evaluated their methods on the Kamitani's dataset. However, this discrepancy hinders the real progress in this line of research. We mainly compare our method with End2EndDIR and SSReconstClass (the improved version of ssfmri2im) because they provide all reconstructed images or the full code to run on this dataset. We cannot find the full published codes or full reconstructed images of the methods - Fang, et al. [20] and Ren, et al. [39], and thus exclude them from the comparisons. In addition, our goal is to perform pixel-level image reconstruction, and therefore we do not compare with some methods such as IC-GAN [25], which focuses on the generation task using semantic consistency.

Fig. 3 shows all the results. Four main conclusions can be drawn. First, our FDGen achieves the best overall performance on the three subjects. Specifically, our FDGen achieves the 14 best results out of a total of 15 metrics compared to End2EndDIR. When reconstructing smaller images, FDGen also achieves significantly better performance than SSReconstClass in 12 out of 15 evaluated metrics. Second, the proposed FDGen achieves large improvements in LPIPS-s scores on all three subjects. For example, FDGen reduces the LPIPS-s score of Subject 1 by 10.5% and 43.4% as compared to End2EndDIR and SSReconstClass respectively. Third, the performance of smaller image reconstruction looks worse in SSIM-s, but significantly better in terms of SSIM-c and LPIPS-s scores. Fourth, all methods show different results across subjects. We speculate that this is due to the fact that fMRI signals used for decoding contain unpredictable noise. In summary, these results confirm that natural image reconstruction from human brain activity is very challenging, and our FDGen achieves better performance compared to previous methods.

In addition, Fig. 4 shows qualitative comparisons of different methods. ssfmri2im can reconstruct coarse layouts of objects. End2EndDIR captures clearer object layouts and shapes. Our FDGen generally produces more accurate object shapes, layouts, color, and texture details.

3.2. *Testing the effectiveness of modules by ablation studies*

To test the effectiveness of FSM and FIM, we also examined the performance of five different variants of our model. The first is **VIM + DC (G1)**. This generator does not use any of the modules proposed in this

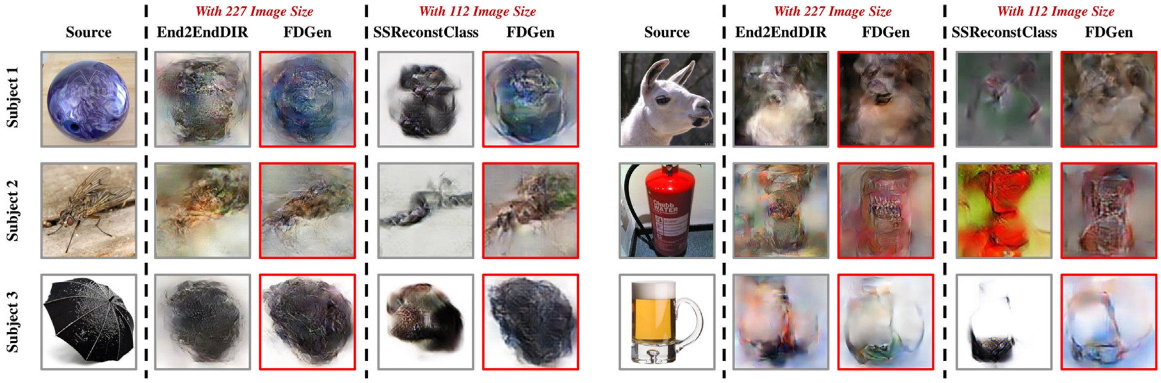

Fig. 4. Qualitative comparisons on the Shen's dataset [16] with two state-of-the-art methods – End2EndDIR [16] and SSReconstClass [19]. For a comprehensive comparison, we include two variants of FDGen for different image sizes, shown as FDGen 112 and FDGen 227. Our FDGen can produce better object shapes, global layouts, and texture details, leading to more reasonable results. Best viewed in the color form.

Table 1
Ablation studies on the Shen's dataset. For evaluation metrics, ↑ indicates a higher value is better, while ↓ means a lower value is better. The best and the second best are highlighted using red and blue. VIM: the convention vector input-based module; DC: only a deconvolutional and a convolutional layer in each intermediate generation block; FIM: function-based input module; FSM-A: selecting features using raw brain activity; FSM-F: selecting features using function-wise embedded features. Best viewed in color format.

Method	Metric: A: PCC-c (↑), B: SSIM-c (↑), C: PCC-s (↑), D: SSIM-s (↑), E: LPIPS-s (↓)														
	Subject 1					Subject 2					Subject 3				
G1. VIM + DC	0.799	0.645	0.311	0.260	0.617	0.769	0.631	0.291	0.269	0.616	0.802	0.653	0.306	0.264	0.605
G2. FIM + DC	0.808	0.654	0.329	0.284	0.590	0.779	0.640	0.295	0.283	0.601	0.809	0.663	0.294	0.286	0.602
G3. VIM + FSM-A	0.819	0.644	0.314	0.263	0.607	0.782	0.633	0.287	0.278	0.609	0.804	0.643	0.314	0.275	0.603
G4. VIM + FSM-F	0.804	0.651	0.320	0.274	0.590	0.780	0.637	0.281	0.277	0.602	0.818	0.654	0.313	0.271	0.599
G5. FIM + FSM (FDGen)	0.822	0.678	0.332	0.282	0.583	0.784	0.644	0.299	0.288	0.597	0.800	0.673	0.296	0.288	0.587
G5 + PatchD [36]	0.818	0.675	0.332	0.283	0.585	0.770	0.650	0.288	0.284	0.592	0.805	0.669	0.311	0.279	0.588
G5 + UNetD [67]	0.820	0.687	0.327	0.290	0.596	0.789	0.654	0.282	0.293	0.604	0.829	0.682	0.302	0.278	0.604

paper. It processes the brain activity input using the conventional vector-based input module (VIM, Fig. 1c). For subsequent steps, we remove all components in the red dashed box in Fig. 1a except a deconvolution layer for feature upsampling and a convolutional layer for feature refinement. For simplicity, we refer to this combination as DC. The second is **FIM + DC (G2)**. In this variant, we directly replace the VIM with the proposed FIM and combine it with DC for the generation process. The third is **VIM + FSM-A (G3)**. This generator only adds the proposed FSM to control the generation process, and still uses the VIM to process the input brain activity. The FSM in this variant directly selects useful cues from the raw brain activity without using function-wise embedding (2). We refer to this variant as FSM-A. The fourth is **VIM + FSM-F (G4)**. This generator is similar to G3, but selects important information from the features extracted from the function-wise embedding layers. The fifth is full **FDGen (G5)**. This is our full FDGen, consisting of FIM and FSM.

The results obtained by the different generators are shown in Table 1. Clearly, the full FDGen (G5) equipped with the two proposed modules obtains the best performance in almost all test metrics across all subjects. Our G5 achieves 12 best results and 1 second best in a total of 15 evaluations (5 metrics by 3 subjects).

To further verify our method, we show four images processed by different generators on Subject 1 in Fig. 5. First, G2 with the proposed FIM consistently captures more details than G1. For example, G2 obtains a better foreground shape of the crab (*i.e.*, the 2nd and 3rd columns). Second, using our FSM module, G4 also captures more texture details than G1 (see the wine glass). Third, G5 uses the two proposed modules and achieves the best reconstruction, as shown in the last column. Similar improvements are highlighted in the colored boxes. Based on these results, we conclude that (1) FIM helps to capture the global layouts of the target images (G2 *vs.* G1), and (2) FIM+FSM (full FDGen) further improves the reconstruction performance (G5 *vs.* all others).

3.3. *Evolution of reconstructed image during model training*

In Fig. 6, we show how the reconstructed images evolve as training proceeds on Subject 1. These results are obtained by testing our model on the test set at different training epochs. We find that FDGen's reconstructed images at early epochs indeed contain very coarse positions, shapes, and low-frequency layouts of the target objects. During the training process, our model gradually captures information from global layouts, such as positions and shapes, to local textures. As a result, our model eventually produces more reasonable results.

3.4. *FDGen is also robust to various types of discriminators*

So far, we have shown that our generator with a standard image-level discriminator achieves good results as compared to the baseline generator. Here, we further test whether our generator is sensitive to the structure of the discriminator. For this purpose, we examined two additional discriminators. The first one is a patch-level discriminator [31], which distinguishes patches as real or fake. The other is a U-shape based discriminator for pixel-level discrimination [40]. We refer to them in Table 1 as PatchD and UNetD, respectively (Fig. 7, reconstruction results). Note that we do not fine-tune the hyperparameters (*e.g.*, batch size, learning rate) when comparing the discriminators. Therefore, the results presented in Table 1 show that our FDGen is insensitive to the discriminator used.

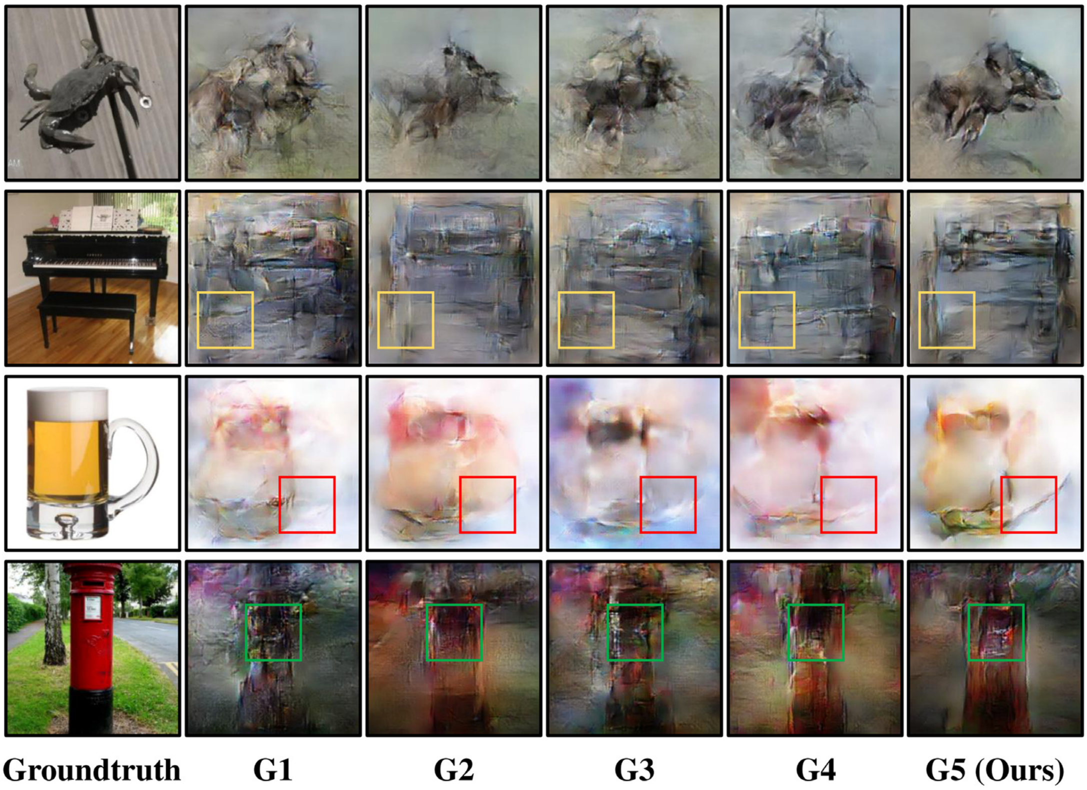

Fig. 5. Qualitative comparisons of the five generators on subject1 in the Shen's dataset. Detail structures of these generators are presented in the main texts. Our full generator G5, combined with the proposed FIM and FSM, produces more details (*e.g.*, edges, textures) and the generated images look more similar to the groundtruth images (1st column). Best views in the color form.

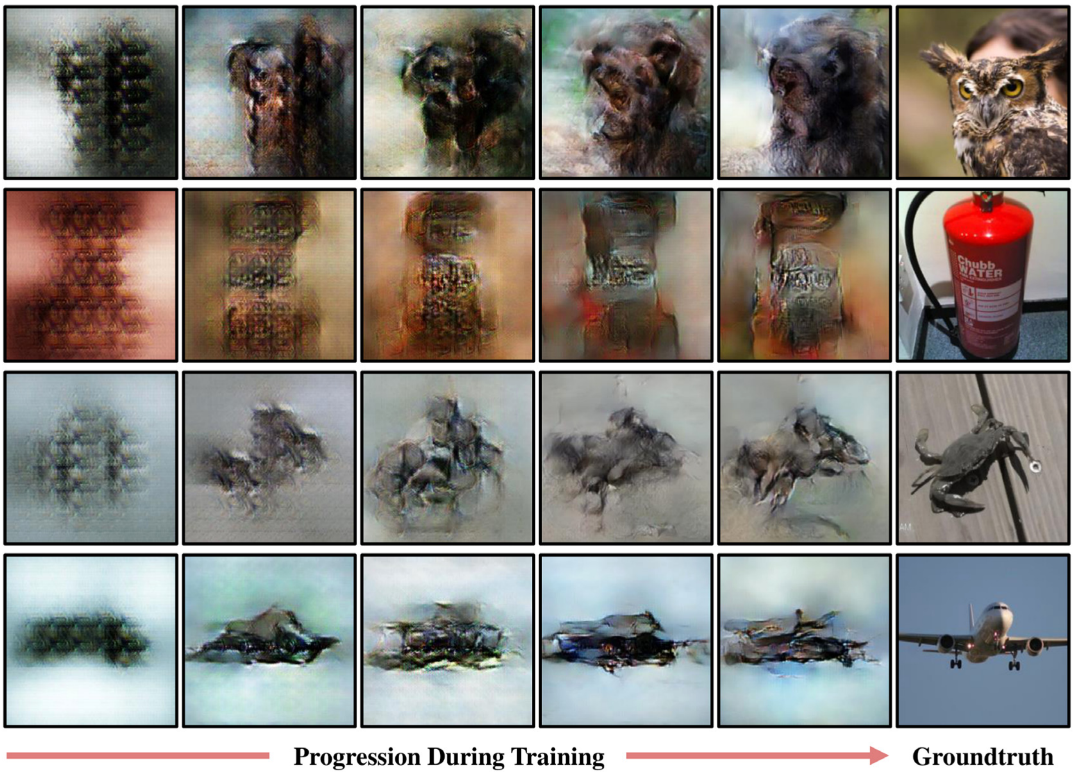

Fig. 6. Four example reconstructed images of Subject 1 during the training process. The generator is our full FDGen (G5) with FIM and FSM modules. All reconstructed images are obtained by testing the model on the test set at different training epochs. Here, the outputs from our model at early epochs contain coarse locations and shapes. Those results are gradually refined when more iterations are achieved.

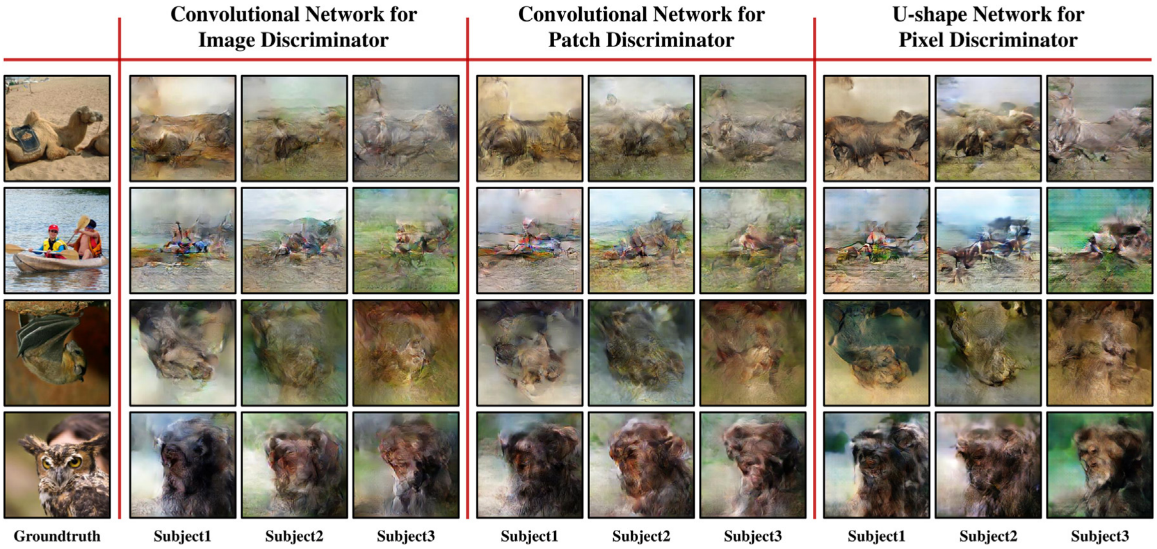
Convolutional Network for Image Discriminator
Convolutional Network for Patch Discriminator
U-shape Network for Pixel Discriminator
Groundtruth
Subject1
Subject2
Subject3
Subject1
Subject2
Subject3
Subject1
Subject2
Subject3

Fig. 7. Four example reconstructions on all subjects in the Shen's dataset. These results are obtained by the proposed generator trained with different discriminators. The discriminator for image discrimination is the default one in our experiment. Overall, our generator is insensitive to the employed discriminator. Best views in the color form.

4. Discussion

In this paper, we propose a novel generator for reconstructing natural images from brain activity measured by fMRI. Our generator contains two key building blocks: a function-based input module and a function-selection module. The function-based input module assumes that different brain regions represent different functions and thus their activity should be processed differently. Our function-selection module automatically learns the weights of the features and refines the features to control the reconstruction process. Extensive experiments on a popular fMRI-based benchmark dataset show that our generator can capture more shape and texture details than previous methods, resulting in more robust reconstruction performance.

4.1. Natural image reconstruction

Early image reconstruction studies typically use simple image patterns, such as gray-scale checkerboard stimuli, or images in specific domains, such as handwritten digits or human faces. Thirion, et al. [11] pioneered an "inverse retinotopic" study, and built a linear model to inversely estimate visual checkerboard stimuli from retinotopic responses in human visual cortex. This is in contrast to conventional retinotopic studies which compute brain responses based on checkerboard stimuli. Following this work, many studies have been proposed to improve reconstruction performance, including linear encoding models [12], and Bayesian models [41,42]. More recently, deep generative networks [14,43] have achieved good reconstruction results in such simple pattern reconstruction tasks [14,43].

Reconstruction of natural images from brain activity has attracted much research attention. There are two main categories of methods to perform reconstruction: (1) designing a generative network, which is usually trained from scratch [16,44]; (2) adapting a pre-trained image generator [15,17,45]. In the first category, researchers always focus on both network design and training strategies. In the second category, researchers often use different methods to explore existing pre-trained models. For example, both Gu, et al. [45] and Takagi and Nishimoto [15] use object categories for training. In addition, due to the different input distribution (*i.e.*, random noise input *vs*. brain voxels) between the pre-training step and the adjustment phase, most methods in the second category tend to synthesize image content rather than to reconstruct images at the pixel level. This may be suboptimal for some real-world applications, such as patient diagnosis. Therefore, we argue that pixel-level constraint-free (*e.g.*, no strong prior knowledge such as semantic category and object locations) natural image reconstruction from brain activity is still very challenging. The measurement noise, subjective factors, and insufficient data make it difficult to be accurately modeled by existing methods [46,47]. Our work here follows the first category, which aims to use an end-to-end generator to perform fine-grained image reconstruction at the pixel level. As a result, we only compare our generator with other related end-to-end methods.

Moreover, most existing models do not consider our brain functions as strong priors for image reconstruction. Only a few recent studies have tested the contributions of individual ROIs [15,45]. Our work considers our brain functions and implements two methods based on two well-known neuroscience theories to achieve good reconstruction results.

4.2. Function parcellations as an appropriate prior to process brain activity for reconstruction

Traditional vector-based input modules used in many previous studies [16,17,19,39] treat all activity as a single input vector for reconstruction (Fig. 1C). However, it is well established that neural representations in the human visual cortex are organized as distinct functional modules [28,48]. For example, numerous human imaging studies have shown that early visual areas V1-V3 can be delineated based on retinotopic responses [28]. In addition, electrophysiological studies have shown that neurons in low-level visual areas (*e.g.*, V1, V2, V3) respond preferentially to simple visual features such as edges, curves, and textures [49]. Neurons in mid-level visual areas (*e.g.*, V4) encode mid-level visual features, such as surfaces or figure-ground segregation [50]. Higher-level visual areas are known to represent global or semantic information [51], such as object shape (*e.g.*, lateral occipital complex area, LOC), faces (*e.g.*, fusiform face area, FFA), and scene (*e.g.*, parahippocampal place area, PPA [32]). In summary, image reconstruction should take into account the functional organization of the human visual cortex. An identical transformation of activity in all brain areas may conflate the contributions of their different functions to the reconstruction, leading to suboptimal performance.

4.3. Attention as a function-selection mechanism to improve generation process

A simple way to map features onto images is to use multiple deconvolutional layers. However, we argue that, in addition to the intrinsic functional parcellations of the human visual cortex, our brain can selectively enhance the subset of information that is important for image reconstruction, as shown in Reynolds and Chelazzi [52]. This is known in neuroscience as attentional selection, one of the most important selection mechanisms that prioritizes task-irrelevant information and attenuates irrelevant signals [53]. For example, goal-directed attention can significantly enhance neural responses to either a spatial location [54] or a feature [55].

The proposed module is related to Hu, et al. [56] but uses a different design. Hu, et al. [56] use features to generate weights and then in turn apply weights to those features. Our FSM can be viewed as a cross-modal attention module that uses global contexts of current features to control the importance of activity in each brain area and then uses the function-wise importance weights to refine the original features. Details of our FSM are shown in Fig. 1a. The generated weights are conditionally dependent on the features of each block, while the output of each block also accepts different cues from different scales of brain activity. We have shown that this input-based adaptive selection process improves the performance of our generator in our experiments.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

CRediT authorship contribution statement

Lingxiao Yang: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Writing – review & editing, Visualization, Writing – original draft. **Hui Zhen:** Validation. **Le Li:** Validation, Writing – review & editing. **Yuanning Li:** Resources, Validation, Writing – review & editing. **Han Zhang:** Resources, Validation, Writing – review & editing. **Ru-Yuan Zhang:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing, Writing – review & editing.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (32100901), Shanghai Pujiang Program (21PJ1407800), Natural Science Foundation of Shanghai (21ZR1434700), Key Laboratory of Brain-Machine Intelligence for Information Behavior Ministry of Education (2023JYBKFKT005) to R.-Y.Z.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2023.08.010.

References

[1] J.J. Shih, D.J. Krusienski, J.R. Wolpaw, Brain-computer interfaces in medicine, in: Proceedings of the Mayo Clin Proc, Elsevier, 2012.
[2] F. Lotte, M. Congedo, A. Lecuyer, et al., A review of classification algorithms for EEG-based brain-computer interfaces, J. Neural Eng. 4 (2007) R1–R13.
[3] R. Sitaram, N. Weiskopf, A. Caria, et al., fMRI brain-computer interfaces, IEEE Signal Process. Mag. 25 (2007) 95–106.
[4] F. Tong, M.S. Pratte, Decoding patterns of human brain activity, Annu. Rev. Psychol. 63 (2012) 483–509.
[5] J.V. Haxby, A.C. Connolly, J.S. Guntupalli, Decoding neural representational spaces using multivariate pattern analysis, Annu. Rev. Neurosci. 37 (2014) 435–456.
[6] J.C. Bore, P. Li, L. Jiang, et al., A long short-term memory network for sparse spatiotemporal EEG source imaging, IEEE Trans. Med. Imaging 40 (2021) 3787–3800.
[7] R.M. Cichy, D. Pantazis, A. Oliva, Resolving human object recognition in space and time, Nat. Neurosci. 17 (2014) 455–462.
[8] Y. Kamitani, F. Tong, Decoding seen and attended motion directions from activity in the human visual cortex, Curr. Biol. 16 (2006) 1096–1102.
[9] V. Michel, A. Gramfort, G. Varoquaux, et al., Total variation regularization for fMRI-based prediction of behavior, IEEE Trans. Med. Imaging 30 (2011) 1328–1340.
[10] R. VanRullen, L. Reddy, Reconstructing faces from fMRI patterns using deep generative neural networks, Commun. Biol. 2 (2019) 193.
[11] B. Thirion, E. Duchesnay, E. Hubbard, et al., Inverse retinotopy: inferring the visual content of images from brain activation patterns, Neuroimage 33 (2006) 1104–1116.
[12] S. Schoenmakers, U. Guclu, M. van Gerven, et al., Gaussian mixture models and semantic gating improve reconstructions from human brain activity, Front. Comput. Neurosci. 8 (2014) 173.
[13] H. Lee, B.A. Kuhl, Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex, J. Neurosci. 36 (2016) 6069–6082.
[14] Y. Güçlütürk, U. Güçlü, K. Seeliger, et al., Reconstructing perceived faces from brain activations with deep adversarial neural decoding, in: Proceedings of the Advances in Neural Information Processing Systems, 2017.
[15] Takagi Y., Nishimoto S. High-resolution image reconstruction with latent diffusion models from human brain activity. bioRxiv. (2022) 2022.2011. 2018.517004.
[16] G. Shen, K. Dwivedi, K. Majima, et al., End-to-end deep image reconstruction from human brain activity, Front. Comput. Neurosci. 13 (2019) 21.
[17] G. Shen, T. Horikawa, K. Majima, et al., Deep image reconstruction from human brain activity, PLoS Compt. Biol. 15 (2019) e1006633.
[18] C. Zhang, K. Qiao, L. Wang, et al., Constraint-free natural image reconstruction from fMRI signals based on convolutional neural network, Front. Hum. Neurosci. 12 (2018) 242.
[19] R. Beliy, G. Gaziv, A. Hoogi, et al., From voxels to pixels and back: self-supervision in natural-image reconstruction from fMRI, Adv. Neural Inf. Process. Syst. 32 (2019) 32 (Nips 2019).
[20] T. Fang, Y. Qi, G. Pan, Reconstructing perceptive images from brain activity by shape-semantic gan, in: Proceedings of the Advances in Neural Information Processing Systems, 2020.
[21] M. Mozafari, L. Reddy, R. VanRullen, Reconstructing natural scenes from fMRI patterns using bigbigan, in: Proceedings of the International joint conference on neural networks (IJCNN), IEEE, 2020.
[22] B.L. Edlow, J. Claassen, N.D. Schiff, et al., Recovery from disorders of consciousness: mechanisms, prognosis and emerging therapies, Nat. Rev. Neurol. 17 (2021) 135–156.
[23] P. Bashivan, K. Kar, J.J. DiCarlo, Neural population control via deep image synthesis, Science 364 (2019) eaav9436.
[24] G. Gaziv, R. Beliy, N. Granot, et al., Self-supervised natural image reconstruction and large-scale semantic classification from brain activity, Neuroimage 254 (2022) 119121.
[25] F. Ozcelik, B. Choksi, M. Mozafari, et al., Reconstruction of perceived images from fMRI patterns and semantic brain exploration using instance-conditioned gans, in: Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022.
[26] T.T. Liu, Noise contributions to the fMRI signal: an overview, Neuroimage 143 (2016) 141–151.
[27] T. Naselaris, K.N. Kay, S. Nishimoto, et al., Encoding and decoding in fMRI, Neuroimage 56 (2011) 400–410.
[28] B.A. Wandell, S.O. Dumoulin, A.A. Brewer, Visual field maps in human cortex, Neuron 56 (2007) 366–383.
[29] Mirza M., Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:. 1411.1784 (2014)
[30] S. Reed, Z. Akata, X. Yan, et al., Generative adversarial text to image synthesis, Proceedings of the International Conference on Machine Learning, PMLR, 2016.
[31] P. Isola, J.-Y. Zhu, T. Zhou, et al., Image-to-image translation with conditional adversarial networks, in: Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
[32] A. Stigliani, K.S. Weiner, K. Grill-Spector, Temporal processing capacity in high-level visual cortex is domain specific, J. Neurosci. 35 (2015) 12412–12424.
[33] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. Acm 60 (2017) 84–90.
[34] J. Deng, W. Dong, R. Socher, et al., Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009.
[35] Kingma D.P., Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. (2014)
[36] K. Seeliger, U. Guclu, L. Ambrogioni, et al., Generative adversarial networks for reconstructing natural images from brain activity, Neuroimage 181 (2018) 775–785.
[37] R. Zhang, P. Isola, A.A. Efros, et al., The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
[38] T. Horikawa, Y. Kamitani, Generic decoding of seen and imagined objects using hierarchical visual features, Nat. Commun. 8 (2017) 15037.
[39] Z. Ren, J. Li, X. Xue, et al., Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning, Neuroimage 228 (2021) 117602.
[40] E. Schonfeld, B. Schiele, A. Khoreva, A u-net based discriminator for generative adversarial networks, in: Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2020.
[41] C. Du, C. Du, H. He, Sharing deep generative representation for perceived image reconstruction from human brain activity, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE, 2017.
[42] T. Naselaris, R.J. Prenger, K.N. Kay, et al., Bayesian reconstruction of natural images from human brain activity, Neuron 63 (2009) 902–915.
[43] C. Du, C. Du, L. Huang, et al., Reconstructing perceived images from human brain activities with Bayesian deep multiview learning, IEEE Trans. Neural Netw. Learn. Syst. 30 (2019) 2310–2323.

[44] K. Han, H. Wen, J. Shi, et al., Variational autoencoder: an unsupervised model for encoding and decoding fMRI activity in visual cortex, Neuroimage 198 (2019) 125–136.

[45] Z. Gu, K.W. Jamison, M. Khosla, et al., Neurogen: activation optimized image synthesis for discovery neuroscience, Neuroimage 247 (2022) 118812.

[46] Z.T. Lu, Visualizing the mind's eye: a future perspective on applications of image reconstruction from brain signals to psychiatry, Psychoradiology 3 (2023) kkad922.

[47] C. Zhang, Y.F. Wang, X.J. Jing, J.H. Yan, Brain mechanisms of mental processing: from evoked and spontaneous brain activities to enactive brain activity, Psychoradiology 3 (2023) Kkad010.

[48] B.A. Wandell, Computational neuroimaging of human visual cortex, Annu. Rev. Neurosci. 22 (1999) 145–173.

[49] M. Carandini, J.B. Demb, V. Mante, et al., Do we know what the early visual system does? J. Neurosci. 25 (2005) 10577–10597.

[50] A.W. Roe, L. Chelazzi, C.E. Connor, et al., Toward a unified theory of visual area v4, Neuron 74 (2012) 12–29.

[51] K. Grill-Spector, R. Malach, The human visual cortex, Annu. Rev. Neurosci. 27 (2004) 649–677.

[52] J.H. Reynolds, L. Chelazzi, Attentional modulation of visual processing, Annu. Rev. Neurosci. 27 (2004) 611–647.

[53] M.M. Chun, J.D. Golomb, N.B. Turk-Browne, A taxonomy of external and internal attention, Annu. Rev. Psychol. 62 (2011) 73–101.

[54] A. Martinez, L. Anllo-Vento, M.I. Sereno, et al., Involvement of striate and extrastriate visual cortical areas in spatial attention, Nat. Neurosci. 2 (1999) 364–369.

[55] J.H. Maunsell, S. Treue, Feature-based attention in visual cortex, Trends Neurosci. 29 (2006) 317–322.

[56] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

Lingxiao Yang (Member, IEEE) received the Ph.D. degree from The Hong Kong Polytechnic University, China, in March 2020. He is currently a postdoctoral researcher working at Sun Yat-sen University since August 2020. He has published several papers (>20) in the prestigious computer science journals and conferences. His research interests include computer vision and machine learning with a focus on object recognition, video understanding, and brain-inspired computational models.

Ru-Yuan Zhang received his Ph.D in Brain&Cognitive Sciences from the University of Rochester. His main research interests include computational mechanisms of visual processing, similarities and differences between deep learning models and human brain processing, computational psychiatry based on psychiatric disorders such as schizophrenia and childhood autism, and analysis and modeling of large-scale brain imaging data for psychiatric disorders. He is currently funded by the National Natural Science Foundation of China, Shanghai Pujiang Program, and Shanghai Natural Science Foundation.