Conditional Generative Neural Decoding with Structured CNN Feature Prediction

Changde Du,[1,2,3] **Changying Du,**[4] **Lijie Huang,**[1] **Huiguang He**[1,2,5,*]
[1]Research Center for Brain-Inspired Intelligence & NLPR, CASIA, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Huawei Cloud BU EI Innovation Lab, China
[4]Huawei Noah's Ark Lab, Beijing 100085, China
[5]Center for Excellence in Brain Science and Intelligence Technology, CAS, Beijing, China
{duchangde, ducyatict}@gmail.com, {lijie.huang, huiguang.he}@ia.ac.cn

Abstract

Decoding visual contents from human brain activity is a challenging task with great scientific value. Two main facts that hinder existing methods from producing satisfactory results are 1) typically small paired training data; 2) underexploitation of the structural information underlying the data. In this paper, we present a novel conditional deep generative neural decoding approach with structured intermediate feature prediction. Specifically, our approach first decodes the brain activity to the multilayer intermediate features of a pretrained convolutional neural network (CNN) with a structured multi-output regression (SMR) model, and then inverts the decoded CNN features to the visual images with an introspective conditional generation (ICG) model. The proposed SMR model can simultaneously leverage the covariance structures underlying the brain activities, the CNN features and the prediction tasks to improve the decoding accuracy and interpretability. Further, our ICG model can 1) leverage abundant unpaired images to augment the training data; 2) self-evaluate the quality of its conditionally generated images; and 3) adversarially improve itself without extra discriminator. Experimental results show that our approach yields state-of-the-art visual reconstructions from brain activities.

Introduction

Brain-sensing signals usually convey rich information about the external stimuli. Effectively decoding these signals could lead to new insights into brain function and boost the development of brain-computer interfaces (BCIs). A challenging task in this filed is visual information decoding, which aims to classify (Haxby et al. 2001; Kamitani and Tong 2005), identify (Kay et al. 2008; Horikawa and Kamitani 2017) or reconstruct (Miyawaki et al. 2008; Naselaris et al. 2009) the perceived visual stimuli from the evoked brain activities measured by functional magnetic resonance imaging (fMRI). Recently, the hierarchical visual features from convolutional neural networks (CNNs) are found to be helpful for connecting the visual stimuli and brain signals. To overcome the small paired training data issue, researchers have tried to use external database to pretrain the CNNs and then reconstruct human faces (Güçlütürk et al. 2017), handwritten characters (Du et al. 2018a) and natural images (Shen et al. 2019) from the CNN features decoded from brain activities.

In deep neural decoding, the most critical point is how to convert the brain activities into the hierarchical CNN features. Almost all existing methods assume that the units of a CNN are conditionally independent of each other, and hence they fit multiple independent single-output linear regression models to decode the CNN features from the fMRI voxels (Horikawa and Kamitani 2017; Wen et al. 2017). However, the CNN features are often related to each other via some underlying structures, and this structural information may be useful for decoding prediction. Throughout the paper, we refer to the dependencies among the different CNN features as ***output structure***, which can be seen as the covariance structure in the output noise. Also, the fMRI voxels are generally highly correlated, and the correlation can carry relevant information about the stimuli. Exploiting the dependencies among the different voxels not only benefits generalization but also helps us to understand the functional networks in the human brain. We refer to the dependencies among the fMRI voxels as ***input structure***. Finally, treating each single-output regression model as a learning task, we refer to the dependencies among different tasks as ***task structure***, which can be captured under the multi-task learning (by imposing some appropriate constraints on the regression coefficients of different tasks). Traditional neural decoding methods ignore these structural informations when modelling the mappings between multivariate inputs and outputs. This greatly limits their decoding performance and interpretability. It is therefore desirable to simultaneously leverage all these three kinds of structures in deep neural decoding for improved predictions and better interpretability.

On the other hand, how to accurately reconstruct the corresponding image based on the decoded CNN features is also a very important problem. Typically, this problem can be solved by maximum a posteriori (MAP) estimation. For example, one can apply gradient-based optimization to find an optimal image that minimizes the reconstruction error in the CNN feature space with a natural image prior such as total variation (TV) regularizer (Mahendran and Vedaldi 2015) or deep generation network (Nguyen et al. 2016; Shen et al. 2019). Alternatively, one can train a de-convolutional neural network (De-CNN) on a large training set of natural images to minimize the reconstruction error in image space (Doso-

vitskiy and Brox 2016b). Nevertheless, both ways mentioned above have their own drawbacks. The former is very slow and the latter often leads to blurry results, which are not desired. Recent studies have shown that the conditional generative adversarial nets (CGANs) (Mirza and Osindero 2014; Dosovitskiy and Brox 2016a) are promising for generating high-fidelity visual images. However, CGANs still face challenges in training stability and sampling diversity. Conditional variational autoencoders (CVAEs) (Sohn, Lee, and Yan 2015) can overcome the weaknesses of CGANs, but generally yield not that realistic images.

Contributions. In this paper, we present a novel conditional deep generative neural decoding approach with structured intermediate feature prediction (see Figure 1). Specifically, our approach first decodes the brain activity to the multilayer intermediate features of a pretrained convolutional neural network (CNN) with a structured multi-output regression (SMR) model, and then inverts the decoded CNN features to the visual images with an introspective conditional generation (ICG) model. Our main contributions can be summarized as follows.

- Employing a matrix-variate Gaussian prior, we develop a Bayesian structured multi-output regression (SMR) model which can simultaneously leverage the covariance structures underlying the fMRI voxels, CNN units and prediction tasks to improve the neural decoding performance.
- We build an introspective conditional generation (ICG) model, which can 1) leverage abundant unpaired images to augment the training data; 2) self-evaluate the quality of its conditionally generated images; and 3) adversarially improve itself without extra discriminator.
- We collected a new fMRI dataset, which can be used for validating the performance of neural decoding models. Our fMRI data will be shared online.
- Experimental results show that our approach yields state-of-the-art visual reconstructions from brain activities.

Related Work

DNN-based neural decoding. Many neural decoding methods used deep neural network (DNN) features for perceived image reconstruction in recent years (Wen et al. 2017; Du, Du, and He 2017; Shen et al. 2019). For example, (Wen et al. 2017) first used linear regression to convert brain activities into CNN features, and then they trained the deconvolutional neural network to perform image reconstruction. (Shen et al. 2019) also used linear regression to convert brain activities into CNN features, and then they applied gradient-based optimization to find an optimal image which minimizes the reconstruction loss of the decoded CNN features. The above methods assume that the units of a CNN are conditionally independent of each other, and hence they fit multiple independent single-output linear regression models to decode the CNN features from the fMRI voxels. This limits their decoding performance and interpretability. Different from them, we developed a structured multi-output regression model to decode the CNN features, and then we use a conditional deep generative models (DGMs) to reconstruct the visual images.

Multi-output regression. Most of the exsiting multi-output regression models are restrictive in the sense that 1) they usually exploit partial structures (*input structure*, *output structure* or *task structure*, but not all), and 2) their prior assumptions about such structures are too strong which may not always be appropriate. For example, the multivariate regression with covariance estimation (MRCE) model (Rothman, Levina, and Zhu 2010) only considered the output structure and therefore fails to account for the input and task structures. Further, (Rai, Kumar, and Daume 2012) proposed an extension of the MRCE model by simultaneously leveraging the output and task structures, but this model still cannot capture the correlation among the inputs. More recently, (Zhao et al. 2017) proposed a multi-task learning model by explicitly modeling the input and task structures, but this model ignores the output structure which is important in multi-output regression problems. In contrast, our SMR is a more comprehensive extension over the above multi-output regression models. It simultaneously leverages all three kinds of structures which are learned from the data automatically.

Generative neural decoding. Recently, there has been research interest in applying DGMs to neural decoding (Güçlütürk et al. 2017; Seeliger et al. 2018; Du et al. 2018a; 2018b; Han et al. 2019; Du, Du, and He 2019). For example, (Du et al. 2018a) proposed a multi-view DGM, in which they first used the sparse linear model to map the brain activities to the latent representation of the VAE, and then used the VAE decoder to reconstruct the image. Furthermore, (Güçlütürk et al. 2017) proposed an adversarial neural decoding method by combining probabilistic inference with the GAN idea. (Du et al. 2018b) developed a multi-view adversarially learned inference model, which formulates the neural decoding problem as a cross-view retrieval task. However, these methods still face challenges in improving neural decoding quality or training stability. Our ICG model is motivated by the success of introspective adversarial learning in image generation (Huang et al. 2018), which combines the advantages of both VAEs and GANs.

Methodology

In the neural decoding dataset, we assume $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]^\top \in \mathbb{R}^{N\times M}$ and $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^\top \in \mathbb{R}^{N\times D}$ denote the matrices of N training visual images and the evoked fMRI activities, respectively. Here, M and D denote the dimensions of $\mathbf{Y}$ and $\mathbf{X}$, respectively. For any image $\mathbf{y}_n$, its hierarchical visual features can be obtained by forward propagating it through a pretrained CNN model (e.g., AlexNet (Krizhevsky, Sutskever, and Hinton 2012) considered in this paper). Here, we use $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_N]^\top \in \mathbb{R}^{N\times K}$ to denote the intermediate CNN features of all training images $\mathbf{Y}$, where K denotes the number of units in that CNN. The proposed approach involves two cascaded stages: *Voxel2Unit* and *Unit2Pixel* (see Figure 1). Below, we will introduce them respectively.

Voxel2Unit: structured multi-output regression

In this stage, we develop a structured multi-output regression (SMR) model to simultaneously leverage the covariance

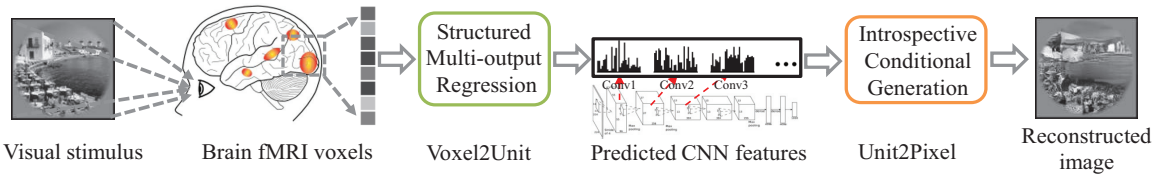


Figure 1: An overview of the proposed neural decoding framework. It involves two cascaded stages, 1) *Voxel2Unit*: decoding the CNN features from fMRI activity and 2) *Unit2Pixel*: reconstructing the perceived image using the decoded CNN features.

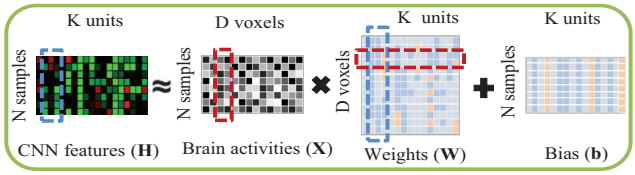


Figure 2: Voxel2Unit: Structured multi-output regression. The red and blue dashed rectangles represent the possible dependencies between the inputs and the outputs, respectively.

structures underlying the fMRI voxels, the CNN features and the prediction tasks (each single-output regression model is regarded as a task) to improve the neural decoding accuracy and interpretability. Specifically, our goal is to learn the functional relationship between the inputs $\mathbf{x}_n \in \mathbb{R}^D$ (fMRI voxels) and the outputs $\mathbf{h}_n \in \mathbb{R}^K$ (CNN features) (see Figure 2), i.e.,

$$\mathbf{h}_n = \mathbf{W}^\top \mathbf{x}_n + \mathbf{b} + \boldsymbol{\epsilon}_n \qquad \forall n = 1, ..., N. \tag{1}$$

Here $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K] \in \mathbb{R}^{D\times K}$ denotes the weight matrix where its column $\mathbf{w}_k \in \mathbb{R}^D$ denotes the regression coefficient of the k-th output and its row $\mathbf{w}^d \in \mathbb{R}^K$ denotes the corresponding coefficient of the d-th input. $\mathbf{b} = [b_1, ..., b_K]^\top \in \mathbb{R}^K$ is a vector of bias terms for the K outputs, and $\boldsymbol{\epsilon}_n = [\epsilon_{n1}, ..., \epsilon_{nK}]^\top \in \mathbb{R}^K$ is a vector consisting of the Gaussian noise for each of the K outputs.

Prior with input and task structures. To take into account both the covariance among the fMRI voxels and the covariance among the prediction tasks, we assume the following (unnormalized) structured prior distribution on $\mathbf{W}$:

$$p(\mathbf{W}) = \Big(\prod_{k=1}^{K} \mathcal{N}(\mathbf{w}_k|\mathbf{0}, \mathbf{I}_D)\Big)\mathcal{MN}(W|\mathbf{0}_{D\times K}, \boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c), \tag{2}$$

where $\mathbf{I}_D$ is a $D \times D$ identity matrix, $\mathbf{0} \in \mathbb{R}^D$ is a D-dimensional vector of all 0s, $\mathcal{MN}(\mathbf{M}, \mathbf{A}, \mathbf{B})$ denotes a matrix-variate Gaussian distribution (Gupta and Nagar 2018) with mean $\mathbf{M} \in \mathbb{R}^{D\times K}$, row covariance matrix $\mathbf{A} \in \mathbb{R}^{D\times D}$ and column covariance matrix $\mathbf{B} \in \mathbb{R}^{K\times K}$. In this structured prior, the $\mathcal{N}(\mathbf{w}_k|\mathbf{0}, \mathbf{I}_D)$ factors regularize the weight vectors $\mathbf{w}_k$ *individually*, and the $\mathcal{MN}(W|\mathbf{0}_{D\times K}, \boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c)$ factor *couples* the D rows of $\mathbf{W}$ by the covariance matrix $\boldsymbol{\Sigma}_r$, and the K columns of $\mathbf{W}$ by the covariance matrix $\boldsymbol{\Sigma}_c$. In other words, the *input structure* and the *task structure* can be discovered by learning $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_c$, respectively.

Likelihood with output structure. The above structured prior can characterize a part of the correlations among the K outputs, but it may be incomplete, due to the limited expressive power of linear predictors. To take into account the potentially residual structural information among the K outputs (i.e., *output structure*) that is not explained by the task structure, we assume a full covariance matrix $\boldsymbol{\Omega} \in \mathbb{R}^{K\times K}$ on the output Gaussian noise distribution. For a set of N i.i.d. observations, the likelihood can be written as:

$$p(\mathbf{H}|\mathbf{X}, \mathbf{W}, \mathbf{b}) = \prod_{n=1}^{N} \mathcal{N}\big(\mathbf{h}_n|\mathbf{W}^\top \mathbf{x}_n + \mathbf{b}, \boldsymbol{\Omega}\big). \tag{3}$$

Note that most previous neural decoding methods (Horikawa and Kamitani 2017; Wen et al. 2017; Shen et al. 2019) can not capture the *output structure* information among the K outputs, since they assume the output Gaussian noise distribution has a diagonal covariance (i.e., $\boldsymbol{\Omega} = \mathbf{I}$).

Given the prior distribution over $\mathbf{W}$ and the likelihood function, we can write down the posterior distribution of $\mathbf{W}$:

$$p(\mathbf{W}|\mathbf{X}, \mathbf{H}, \mathbf{b}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c) \propto \Big(\prod_{n=1}^{N} \mathcal{N}\big(\mathbf{h}_n|\mathbf{W}^\top \mathbf{x}_n + \mathbf{b}, \boldsymbol{\Omega}\big)\Big) \Big(\prod_{k=1}^{K} \mathcal{N}(\mathbf{w}_k|\mathbf{0}, \mathbf{I}_D)\Big) \cdot \mathcal{MN}(W|\mathbf{0}_{D\times K}, \boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c). \tag{4}$$

Taking the log of the Eq. (4) and simplifying the resulting expression (ignoring the constants), we can solve $\mathbf{W}$ by maximum a posteriori (MAP) estimation. Specifically, the negative log-posterior of $\mathbf{W}$ can be written as:

$$\begin{aligned}\mathcal{J} = &\operatorname{tr}\big((\mathbf{H} - \mathbf{X}\mathbf{W} - \mathbf{1}\mathbf{b}^\top)\boldsymbol{\Omega}^{-1}(\mathbf{H} - \mathbf{X}\mathbf{W} - \mathbf{1}\mathbf{b}^\top)^\top\big) \\ &- N \log|\boldsymbol{\Omega}^{-1}| + \lambda \operatorname{tr}(\mathbf{W}\mathbf{W}^\top) + \lambda_1 \operatorname{tr}(\boldsymbol{\Sigma}_r^{-1}\mathbf{W}\boldsymbol{\Sigma}_c^{-1}\mathbf{W}^\top) \\ &- K \log|\boldsymbol{\Sigma}_r^{-1}| - D \log|\boldsymbol{\Sigma}_c^{-1}|,\end{aligned} \tag{5}$$

where $\operatorname{tr}(\cdot)$ denotes matrix trace, $\mathbf{1}$ denotes a $N \times 1$ vector of all 1s, and $|\cdot|$ denotes matrix determinant. Here, λ and λ_1 are the introduced regularization hyperparameters, which control the trade-off between data-fit and model complexity. Note that the $\operatorname{tr}(\boldsymbol{\Sigma}_r^{-1}\mathbf{W}\boldsymbol{\Sigma}_c^{-1}\mathbf{W}^\top)$ term captures the dependencies among the rows of $\mathbf{W}$ by learning the input inverse covariance matrix $\boldsymbol{\Sigma}_r^{-1}$, and the dependencies among the columns of $\mathbf{W}$ by learning the task inverse covariance matrix $\boldsymbol{\Sigma}_c^{-1}$.

Sparse covariance selection. The inverse covariance matrices $\boldsymbol{\Omega}^{-1}$, $\boldsymbol{\Sigma}_r^{-1}$ and $\boldsymbol{\Sigma}_c^{-1}$ will be learned from the data. Sparsity on these parameters is appealing for two reasons: 1) sparsity leads to improved robust estimates of $\boldsymbol{\Omega}^{-1}$, $\boldsymbol{\Sigma}_r^{-1}$

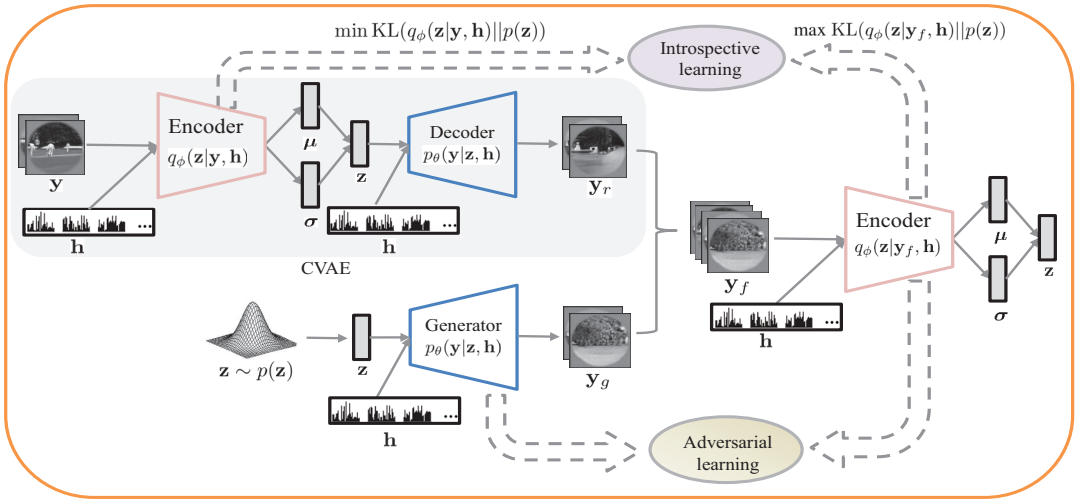


Figure 3: Unit2Pixel: Introspective conditional generation. Weight sharing between the decoder and the generator, and similarly for the two encoders. In test phase, we use the generator $p_\theta(\mathbf{y}|\mathbf{z},\mathbf{h})$ to obtain image reconstructions, where $\mathbf{z} \sim p(\mathbf{z})$ and $\mathbf{h}$ is the decoded CNN features.

and $\mathbf{\Sigma}_c^{-1}$ (Friedman, Hastie, and Tibshirani 2008); 2) sparsity supports the notion that the dependencies among inputs/outputs/tasks tend to be sparse – not all pairs of voxels/units/tasks are related. For example, when $\mathbf{\Sigma}_r^{-1}$ is sparse, a zero entry in it indicates no direct interaction between the two corresponding voxels in the multi-output regression. Similar explanations are in $\mathbf{\Sigma}_c^{-1}$ and $\mathbf{\Omega}^{-1}$. Therefore, we impose sparsity constraints on $\mathbf{\Omega}^{-1}$, $\mathbf{\Sigma}_r^{-1}$ and $\mathbf{\Sigma}_c^{-1}$ via the ℓ_1 penalty. Let $\Theta = \{\mathbf{W}, \mathbf{b}, \mathbf{\Omega}^{-1}, \mathbf{\Sigma}_r^{-1}, \mathbf{\Sigma}_c^{-1}\}$, then the regularized objective function is:

$$\min_{\Theta} \mathcal{J}_s = \mathcal{J} + \lambda_2 ||\mathbf{\Omega}^{-1}||_1 + \lambda_3 (||\mathbf{\Sigma}_r^{-1}||_1 + ||\mathbf{\Sigma}_c^{-1}||_1), \tag{6}$$

where $||\cdot||_1$ is the ℓ_1-norm of a matrix, and $\{\lambda_2, \lambda_3\}$ are the introduced regularization hyperparameters, which control the sparsity of the inverse covariance matrices.

Unit2Pixel: introspective conditional generation

In this stage, our goal is to reconstruct the perceived images using the CNN features decoded from the first stage. Recent studies have shown that it is a promising way to use the conditional DGMs such as conditional variational autoencoders (CVAEs) (Sohn, Lee, and Yan 2015) and conditional generative adversarial nets (CGANs) (Mirza and Osindero 2014; Dosovitskiy and Brox 2016a) to invert the CNN features back to visual images. Treating the output of the SMR model (i.e., the predicted CNN features $\mathbf{H}$) as conditions, the object of CVAE is to minimize the following objective:

$$\mathcal{L}_{\text{CVAE}} = \underbrace{-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})}[\log p_\theta(\mathbf{y}|\mathbf{z},\mathbf{h})]}_{L_{AE}} + \underbrace{\text{KL}(q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})||p(\mathbf{z}))}_{D_{KL}}, \tag{7}$$

where $q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})$ and $p_\theta(\mathbf{y}|\mathbf{z},\mathbf{h})$ are the encoder and the decoder, respectively, with the observable variables $\mathbf{y}$, the latent variables $\mathbf{z}$ and the condition $\mathbf{h}$ (see the gray sub-panel in Figure 3). Here θ and ϕ are the parameters of the decoder and the encoder, respectively. In Eq. (7), the L_{AE} term denotes the reconstruction error, and the D_{KL} term regularizes the encoder by encouraging the approximate posterior $q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})$ to match the prior $p(\mathbf{z})$. Though CVAEs are theoretically elegant and easy to train, it tends to produce blurry images that lack details.

Inspired by previous work (Huang et al. 2018), here we build an introspective conditional generation (ICG) model (see Figure 3), which can self-estimate the differences between conditionally generated and real images and then update itself to produce more realistic images. In model training, the KL divergence term D_{KL} is adversarially optimized along with the reconstruction error term L_{AE}, which increases the difficulty of distinguishing between the real and fake (i.e., reconstructed or generated) images for the encoder. Specifically, the encoder attempts to minimize D_{KL} for real images while maximize it for the fake images. In contrast, the decoder/generator attempts to mislead the encoder by minimizing D_{KL} for the fake images. Formally, we use the following objectives to iteratively optimize the decoder/generator and the encoder until convergence:

$$\hat{\theta} = \arg\min_{\theta}[L_{AE} + \alpha D_{KL}(q_\phi(\mathbf{z}|\mathbf{y}_f,\mathbf{h})||p(\mathbf{z}))], \tag{8}$$

$$\hat{\phi} = \arg\min_{\phi}[L_{AE} + \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})||p(\mathbf{z})) - \alpha D_{KL}(q_\phi(\mathbf{z}|\mathbf{y}_f,\mathbf{h})||p(\mathbf{z}))]. \tag{9}$$

Here $\mathbf{y}_f$ denotes the fake images drawn from $p_\theta(\mathbf{y}|\mathbf{z},\mathbf{h})$, where $\mathbf{z} \sim p(\mathbf{z})$ or $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})$. Note that the L_{AE} term in (8) and (9) builds a bridge between the decoder/generator and the encoder. Intuitively, both cooperative learning and adversarial learning exist between (8) and (9). For the real data points, (8) and (9) are cooperative. Specifically, both (8) and (9) aim to minimize the reconstruction error L_{AE}, and Eq. (9) also tries to minimize the $D_{KL}(q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})||p(\mathbf{z}))$, which

is equivalent to optimize the objective of CVAE in Eq. (7). For the fake data points, (8) and (9) are adversarial. Specifically, (8) aims to minimize the $D_{KL}(q_\phi(\mathbf{z}|\mathbf{y}_f,\mathbf{h})||p(\mathbf{z}))$ term, while (9) aims to maximize it, which forms the adversarial learning like CGAN. α and β are user-defined parameters, which are used to balance the impact of CVAE and CGAN. For example, when $\alpha = 0, \beta = 1$ the proposed method collapses to the standard CVAEs, and when $\alpha = 1, \beta = 0$ the proposed method is equivalent to a regularized CGAN. Compared to most VAE and GAN hybrid models (Larsen et al. 2016; Bao et al. 2017), our ICG model requires no extra discriminator (because the encoder also plays a role of discriminator), which reduces the complexity of the model.

Optimization

Instead of training the above two-stage approach in an end-to-end manner, we learn each stage individually. The motivations are twofold: 1) after the training of the first stage, we can select some CNN units with high decodability for the training of the second stage; 2) end-to-end training needs a large number of paired (stimuli-responses) data, which is usually not satisfied in neural decoding, while we can use the augmented large scale image datasets to train our ICG model individually.

Training SMR model. The objective function $\mathcal{J}_s$ in Eq. (6) is not jointly convex over all variables but is individually convex w.r.t. each variable when others are fixed. Here, we adopt an alternating optimization strategy to learn the proposed SMR model. For example, when $\mathbf{b}, \mathbf{\Omega}^{-1}, \mathbf{\Sigma}_r^{-1}$ and $\mathbf{\Sigma}_c^{-1}$ are fixed, we solve the following subproblem to update $\mathbf{W}$:

$$\min_{\mathbf{W}} \mathcal{L}_{\mathbf{W}} = \mathrm{tr}\big((\mathbf{H}-\mathbf{X}\mathbf{W}-\mathbf{1}\mathbf{b}^\top)\mathbf{\Omega}^{-1}(\mathbf{H}-\mathbf{X}\mathbf{W}-\mathbf{1}\mathbf{b}^\top)^\top\big) + \lambda \mathrm{tr}(\mathbf{W}\mathbf{W}^\top) + \lambda_1 \mathrm{tr}(\mathbf{\Sigma}_r^{-1}\mathbf{W}\mathbf{\Sigma}_c^{-1}\mathbf{W}^\top). \quad (10)$$

Proposition 1. *Eq. (10) can be solved in closed form in* $\mathcal{O}(K^3D^3 + K^2ND^2)$ *time; the optimal solution satisfies:* $\mathrm{vec}(\hat{\mathbf{W}}) = \mathbf{U}^{-1}\mathbf{V}$, *where* $\mathbf{U} = \mathbf{\Omega}^{-1} \otimes (\mathbf{X}^\top\mathbf{X}) + \lambda \mathbf{I}_{KD} + \lambda_1 \mathbf{\Sigma}_c^{-1} \otimes \mathbf{\Sigma}_r^{-1}$ *and* $\mathbf{V} = \mathrm{vec}(\mathbf{X}^\top(\mathbf{H}-\mathbf{1}\mathbf{b}^\top)\mathbf{\Omega}^{-1})$.

The proof is provided in Appendix A. Here $\otimes$ denotes the Kronecker product and $\mathrm{vec}(\mathbf{W})$ denotes the vectorization of $\mathbf{W}$. $\hat{\mathbf{W}}$ can then be obtained simply by reformatting $\mathrm{vec}(\hat{\mathbf{W}})$ into a $D \times K$ matrix. The closed form solution shown above requires us to explicitly form a matrix of size $KD \times KD$. This can be intractable even for moderate K and D. In such cases, we can alternatively use gradient descent method with $\nabla_{\mathbf{W}}\mathcal{L}_{\mathbf{W}} = \mathbf{X}^\top(\mathbf{H}-\mathbf{X}\mathbf{W}-\mathbf{1}\mathbf{b}^\top)\mathbf{\Omega}^{-1} + \lambda\mathbf{W} + \lambda_1\mathbf{\Sigma}_r^{-1}\mathbf{W}\mathbf{\Sigma}_c^{-1}$ to obtain an approximate solution.

The updating rules of the other variables ($\mathbf{b}, \mathbf{\Omega}^{-1}, \mathbf{\Sigma}_r^{-1}$ and $\mathbf{\Sigma}_c^{-1}$) are provied in Appendix B.

Training ICG model. The decoder/generator and the encoder in our ICG model can be jointly trained by optimizing the different objective functions in Eq. (8) and Eq. (9) iteratively. In each iteration, the model parameters (θ or ϕ) can be estimated efficiently in the stochastic gradient variational Bayes (SGVB) (Kingma and Welling 2014) framework. Specifically, we assume the prior $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the encoder $q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})$ is designed to output two individual variables, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, and then we assume the approximated posterior $q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$. In this setting, the KL-divergence term in Eq. (8) and Eq. (9) can be computed as:

$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{y},\mathbf{h})||p(\mathbf{z})) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{d_z}(1+\log(\boldsymbol{\sigma}_{ij}^2) - \boldsymbol{\mu}_{ij}^2 - \boldsymbol{\sigma}_{ij}^2), \quad (11)$$

where d_z is the dimension of the latent variable $\mathbf{z}$. For the reconstruction error L_{AE} in Eq. (8) and Eq. (9), we choose the commonly-used pixel-wise mean squared error:

$$L_{AE}(\mathbf{y},\mathbf{y}_r) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\|\mathbf{y}_{r,ij} - \mathbf{y}_{ij}\|_2^2, \quad (12)$$

where $\mathbf{y}_r$ denotes the reconstructed image.

Experiments

Datasets. Here we briefly introduce the two datasets used in our experiments (and see Appendix C for more details). 1) **Vim-1**: a publicly available fMRI dataset, which contains the blood-oxygen-level dependent (BOLD) responses of two subjects when they are presented with grayscale natural images (Kay et al. 2008). The dataset is partitioned into distinct training and test sets which consist of 1750 and 120 instances, respectively. 2) **FaceBold**: We collected a new fMRI dataset, which comprises grayscale face stimuli and the corresponding BOLD responses of six subjects. In total, 720 faces were presented once for the training set, and 80 faces were presented twice for the test set. For our ICG model, we use the gray scale **ImageNet-1k** (Deng et al. 2009) and **CelebA** (Liu et al. 2015) datasets to augment the training sets of Vim-1 and FaceBold, respectively. The properties of the datasets used in our experiments have been summarized in Table 1.

Table 1: The details of the datasets used in our experiments.

Dataset	Training (N)	Test	Voxels (D)	Units (K)	Pixels
Vim-1	1750	120	8428	438856	128×128
FaceBold	720	80	5000	438856	128×128

Compared methods. In *Voxel2Unit*, we compare our SMR with 1) **SLR**: single-output linear regressions (Horikawa and Kamitani 2017); 2) **BCCA**: Bayesian canonical correlation analysis (Fujiwara, Miyawaki, and Kamitani 2013). In addition, we also study various different configurations of our SMR framework, e.g., 3) **SMR-O**: SMR with only output structure, which corresponds to fixing $\mathbf{\Sigma}_r^{-1} = \mathbf{I}$ and $\mathbf{\Sigma}_c^{-1} = \mathbf{I}$ (Rothman, Levina, and Zhu 2010); 4) **SMR-IT**: SMR with only input and task structures, which corresponds to fixing $\mathbf{\Omega}^{-1} = \mathbf{I}$ (Zhao et al. 2017). In *Unit2Pixel*, we compare our ICG model with 1) **CVAE** (Du et al. 2018a; Sohn, Lee, and Yan 2015); 2) **CGAN**: (Mirza and Osindero 2014; Dosovitskiy and Brox 2016a); 3) **Grad-TV**: gradient-based optimization with a total variation (TV) regularizer (Mahendran and Vedaldi 2015); 4) **De-CNN**: de-convolutional neural network (Dosovitskiy and Brox 2016b).

Parameter settings. For SMR model, we experiment with its two variants. One without the sparsity assumptions on the inverse covariance matrices, and the other with the sparsity assumptions on the inverse covariance matrices. We fix the hyperparameter λ as 0.001 for both cases. For non-sparse case, we fix $\lambda_2 = \lambda_3 = 10^{-6}$, and λ_1 is selected using five-fold cross-validation within the range $[10^{-5}, 10^3]$. For sparse case, we use the same value of λ_1 that was selected for non-sparse case, and only λ_2 and λ_3 are selected by cross-validation. For ICG model, we treat the top 5000 decodable CNN units (according to the rank of each unit's decodability) as condition, and set $\{\alpha = 0.5, \beta = 1\}$ to combine the advantages of both CVAE and CGAN. We set $\{\alpha = 0, \beta = 1\}$ and $\{\alpha = 1, \beta = 0\}$ in ICG to implement CVAE and CGAN, respectively. The latent variable $\mathbf{z}$ is randomly drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{I})$ distribution, with the dimension set to 512 and 256 on the Vim-1 and FaceBold datasets, respectively.

Experimental Results

***Voxel2Unit*: CNN feature decoding.** Here we compare our SMR with non-structured baselines and we also study the ablation of SMR (its various different configurations, SMR-O, SMR-I, etc.). The results on both datasets are summarized in Table 2. Several observations can be drawn as follows. First, by comparing SMR against the baselines, we can find that SMR performs considerably better in all cases. Second, by examining SMR against SLR and BCCA which make no structure assumptions, we can find that SMR always outperform them. This supports our motivation that the covariance structures over the fMRI voxels, CNN features and prediction tasks are important in multi-output regression. Third, SMR shows better performance than its six special cases. This result shows that simultaneously leveraging the multiple covariance structures are also important. Finally, we also note that the sparse cases always perform better or as good as the non-sparse cases on the both datasets, which suggests that explicitly encouraging zero entries in the inverse covariance matrices leads to better estimations of the structures (by avoiding spurious correlations). This can potentially improve the prediction performance.

Table 2: Average normalized mean squared error (NMSE) across 5 random runs (the lower the better).

Method	Vim-1		FaceBold	
	Non-sparse	Sparse	Non-sparse	Sparse
SLR	-	.641 ± .024	-	.724 ± .025
BCCA	-	.693 ± .025	-	.785 ± .026
SMR-T	.582 ± .024	.578 ± .023	.680 ± .025	.672 ± .025
SMR-I	.579 ± .023	.580 ± .025	.676 ± .026	.673 ± .025
SMR-O	.586 ± .023	.578 ± .022	.694 ± .024	.672 ± .024
SMR-OT	.568 ± .024	.564 ± .024	.669 ± .023	.668 ± .023
SMR-IT	.563 ± .023	.560 ± .024	.667 ± .024	.659 ± .025
SMR-IO	.565 ± .024	.562 ± .023	.672 ± .023	.664 ± .023
SMR	**.562** ± .023	**.549** ± .022	**.660** ± .024	**.651** ± .023

Performance layer-by-layer. Figure 4 shows the results of the Pearson correlation coefficient (PCC) between the true CNN feature and the SMR predicted ones. We see that features from deep layers (conv3 to fc8) generalize better than features from shallow layers (conv1, conv2). Intuitively, the use of the hierarchical features with higher prediction accuracy will be good for reconstructing the perceived images.

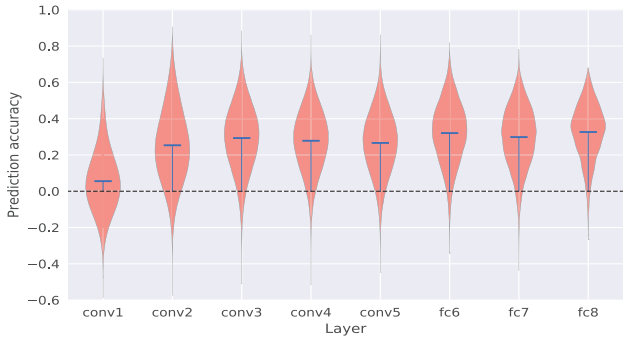


Figure 4: The distributions of decoding accuracy of all individual units in each layer. Blue bars denote mean prediction accuracies averaged across all units.

***Unit2Pixel*: perceived image reconstruction.** Figure 5 shows several representative examples of the test stimuli and their reconstructions based on the CNN features decoded by our SMR model (more results are shown in Appendix D, including image reconstructions based on the CNN features decoded by the baselines). The first row shows the original test stimuli, and the second row shows the reconstructed stimuli from the true CNN features. The second row can be regarded as the upper limit of the reconstruction performance of our ICG model. Because they are the best possible reconstructions that we can expect to achieve with a perfect SMR model that can exactly predict the CNN features from brain activities. The following rows show the reconstructions of all approaches using the CNN features predicted by our SMR. Visual inspection of the reconstructions produced by our ICG model reveals that they match the test stimuli in several key aspects, such as contour, texture and some semantic features. Table 3 shows two evaluation metrics in terms of the ratio of the reconstruction accuracies obtained from the decoded CNN features and the true CNN features. We see that our method achieves better quantitative performance than the competitors on both datasets.

Table 3: Reconstruction accuracy measured by Pearson correlation coefficient (PCC) and structural similarity index (SSIM).

Method	Vim-1		FaceBold	
	PCC	SSIM	PCC	SSIM
Grad-TV	.263 ± .055	.350 ± .039	.374 ± .051	.432 ± .044
De-CNN	.458 ± .044	.545 ± .027	.548 ± .046	.755 ± .037
CVAE	.475 ± .045	.592 ± .030	.577 ± .048	.794 ± .031
CGAN	.493 ± .044	.625 ± .029	.595 ± .045	.831 ± .025
ICG	**.551** ± .044	**.675** ± .024	**.658** ± .042	**.872** ± .025

Covariance structures visualization. We show the learnt sparse inverse covariance matrices in Figure 6. To better visualize, we only present 30 voxels from the brain area V1 and

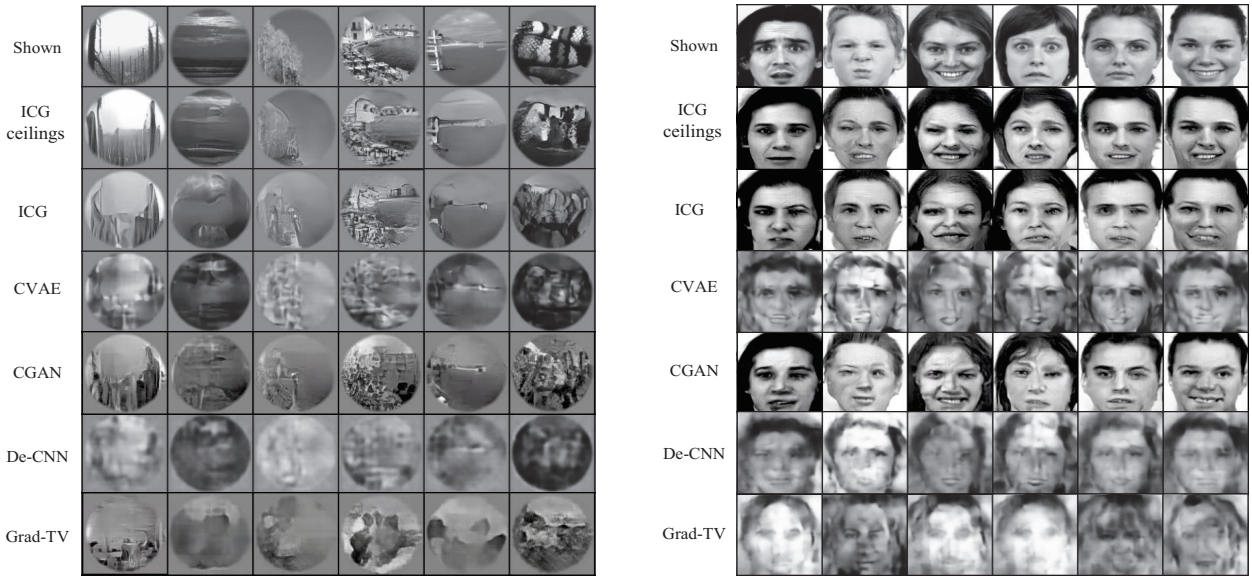
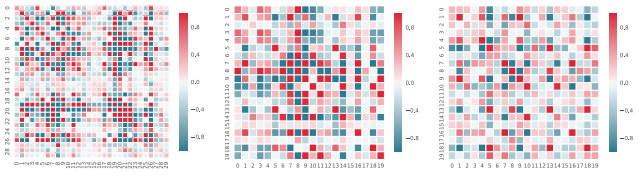


Figure 5: Examples of reconstructed natural images and human faces on the Vim-1 and the FaceBold datasets, respectively. Since only top 5000 decodable CNN units (according to the rank of each unit's decodability) are used as the input, the performance of ICG ceilings maybe not the best. The ceiling performance can be improved if we use more CNN units, but the computing complexity will also increase accordingly.

20 CNN units from the first layer of AlexNet. From Figure 6, we can find some well organized structures, which reflect the mutual dependence among multiple different variables. For example, the voxels $\#20 - \#22$ are highly correlated with each other in Figure 6 (a). We think it is these structural informations that make our SMR model perform better prediction than the baselines.



(a) Input structure $\mathbf{\Sigma}_r^{-1}$ (b) Task structure $\mathbf{\Sigma}_c^{-1}$ (c) Output structure $\mathbf{\Omega}^{-1}$

Figure 6: Visualization of the inverse covariance matrices learned by SMR on the Vim-1 dataset. From Figure 6 (a), we observe that the fMRI voxels are not independent of each other, and the voxels $\#20 - \#22$ are highly correlated with each other.

Convergence study. We also study the convergence properties of our SMR method. Figure 7 shows the plots of the `log` value of the objective function (given by Eq. (6)) and the average MSE with increasing number of iterations on the Vim-1 dataset. The plots show that our alternating optimization procedure converges in roughly 50 iterations.

Conclusion

We have proposed a two-stage neural decoding method to tackle the perceived image reconstruction problem. In the first stage, we developed a structured multi-output regression model, which can simultaneously take into account the covariance structures of the fMRI voxels, the decoding tasks and the CNN features. In the second stage, by combining the maximum likelihood estimation with adversarial learning, we proposed an introspective conditional generation model, which can be trained stably to generate sharper images. Our method can fully explore the structural information underlying the data and can make full use of the large number of image data to improve the neural decoding performance. Experimental results have confirmed the superiority of the proposed method.

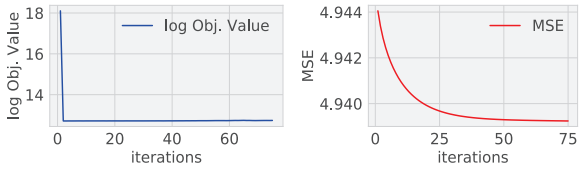


Figure 7: The convergence properties of the proposed SMR model. We see that our SMR method converges in roughly 50 iterations

Leveraging the semi-supervised learning (Du, Du, and He 2019) and cycle consistency learning (Du et al. 2018b) to further improve the neural decoding performance are two promising directions.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61976209, 61906188,

61602449), CAS Scientific Equipment Development Project (YJKYYQ20170050), and Strategic Priority Research Program of CAS (XDB32040200).

References

Bao, J.; Chen, D.; Wen, F.; Li, H.; and Hua, G. 2017. CVAE-GAN: Fine-grained image generation through asymmetric training. In *ICCV*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dosovitskiy, A., and Brox, T. 2016a. Generating images with perceptual similarity metrics based on deep networks. In *NeurIPS*.

Dosovitskiy, A., and Brox, T. 2016b. Inverting visual representations with convolutional networks. In *CVPR*.

Du, C.; Du, C.; Huang, L.; and He, H. 2018a. Reconstructing perceived images from human brain activities with bayesian deep multiview learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Du, C.; Du, C.; Xie, X.; Zhang, C.; and Wang, H. 2018b. Multi-view adversarially learned inference for cross-domain joint distribution matching. In *SIGKDD*.

Du, C.; Du, C.; and He, H. 2017. Sharing deep generative representation for perceived image reconstruction from human brain activity. In *IJCNN*.

Du, C.; Du, C.; and He, H. 2019. Doubly semi-supervised multimodal adversarial learning for classification, generation and retrieval. In *Proc. ICME*.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441.

Fujiwara, Y.; Miyawaki, Y.; and Kamitani, Y. 2013. Modular encoding and decoding models derived from bayesian canonical correlation analysis. *Neural computation* 25(4):979–1005.

Güçlütürk, Y.; Güçlü, U.; Seeliger, K.; Bosch, S.; van Lier, R.; and van Gerven, M. A. 2017. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *NeurIPS*.

Gupta, A. K., and Nagar, D. K. 2018. *Matrix variate distributions*. Chapman and Hall/CRC.

Han, K.; Wen, H.; Shi, J.; Lu, K.-H.; Zhang, Y.; and Liu, Z. 2019. Variational autoencoder: An unsupervised model for modeling and decoding fmri activity in visual cortex. *NeuroImage* 125–136.

Haxby, J. V.; Gobbini, M. I.; Furey, M. L.; Ishai, A.; Schouten, J. L.; and Pietrini, P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293(5539):2425–2430.

Horikawa, T., and Kamitani, Y. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications* 8:15037.

Huang, H.; He, R.; Sun, Z.; Tan, T.; et al. 2018. Introvae: Introspective variational autoencoders for photographic image synthesis. In *NeurIPS*.

Kamitani, Y., and Tong, F. 2005. Decoding the visual and subjective contents of the human brain. *Nature neuroscience* 8(5):679.

Kay, K. N.; Naselaris, T.; Prenger, R. J.; and Gallant, J. L. 2008. Identifying natural images from human brain activity. *Nature* 452(7185):352.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 1097–1105.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2016. Autoencoding beyond pixels using a learned similarity metric. In *ICML*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Mahendran, A., and Vedaldi, A. 2015. Understanding deep image representations by inverting them. In *CVPR*.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Miyawaki, Y.; Uchida, H.; Yamashita, O.; Sato, M. A.; Morito, Y.; Tanabe, H. C.; Sadato, N.; and Kamitani, Y. 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60(5):915–929.

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(6):902–915.

Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; and Clune, J. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *NeurIPS*.

Rai, P.; Kumar, A.; and Daume, H. 2012. Simultaneously leveraging output and task structures for multiple-output regression. In *NeurIPS*.

Rothman, A. J.; Levina, E.; and Zhu, J. 2010. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* 19(4):947–962.

Seeliger, K.; Güçlü, U.; Ambrogioni, L.; Güçlütürk, Y.; and van Gerven, M. 2018. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage* 181:775–785.

Shen, G.; Horikawa, T.; Majima, K.; and Kamitani, Y. 2019. Deep image reconstruction from human brain activity. *PLOS Computational Biology* 15(1):e1006633.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. In *NeurIPS*.

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. *Cerebral Cortex* 1–25.

Zhao, H.; Stretcu, O.; Negrinho, R.; Smola, A.; and Gordon, G. 2017. Efficient multi-task feature and relationship learning. In *NeurIPS*.