# Brain Encoding and Decoding in fMRI with Bidirectional Deep Generative Models

Changde Du [a, b], Jinpeng Li [a, b] , Lijie Huang [a, b] , Huiguang He [a, b, c, *]

[a] *Research Center for Brain-Inspired Intelligence and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, 100190, China*
[b] *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China*
[c] *Center for Excellence in Brain Science and Intelligence Technology CAS, Beijing, China*
[*] *Corresponding author. E-mail: huiguang.he@ia.ac.cn*

**ABSTRACT** Brain encoding and decoding via functional magnetic resonance imaging (fMRI) are two important aspects of visual perception neuroscience. Although previous researchers have made significant advances in brain encoding and decoding models, existing methods still require improvement using advanced machine learning techniques. For example, traditional methods usually build the encoding and decoding models separately, and are prone to overfitting on a small dataset. In fact, effectively unifying the encoding and decoding procedures may allow for more accurate predictions. In this paper, we first review the existing encoding and decoding methods and discuss the potential advantages of a "bidirectional" modeling strategy. Next, we show that there are correspondences between deep neural networks and human visual streams in terms of the architecture and computational rules. Furthermore, deep generative models (e.g., variational autoencoders (VAEs) and generative adversarial networks (GANs)) have produced promising results in studies on brain encoding and decoding. Finally, we propose that the dual learning method, which was originally designed for machine translation tasks, could help to improve the performance of encoding and decoding models by leveraging large-scale unpaired data.

**KEYWORDS** Brain encoding and decoding, fMRI, Deep neural networks, Deep generative models, Dual learning

## 1 Introduction

The relationship between human visual experience and the evoked neural activity is central to the field of computational neuroscience [1, 2]. Brain encoding and decoding via functional magnetic resonance imaging (fMRI) are important in gaining an understanding of the visual perception system [3-5]. An encoding model attempts to predict brain response based on a given visual stimulus [6, 7], whereas a decoding model attempts to predict the corresponding visual stimulus by analyzing a given brain response [8-23]. Brain encoding and decoding (see Figure 1) have become two important ways to promote the development of sensory neuroscience because they provide many insights into brain function.

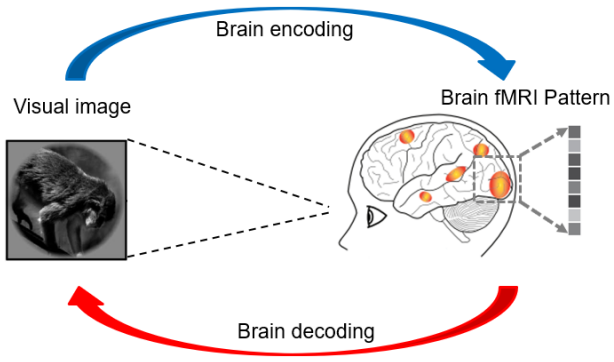


**Figure 1** Brain encoding and decoding in fMRI. The encoding model attempts to predict brain responses based on the presented visual stimuli, while the decoding model attempts to infer the corresponding visual stimuli by analyzing the observed brain responses. In practice, the encoding models and decoding models should not be taken as mutually exclusive. Effective unifying the encoding and decoding procedures may allow for more accurate predictions and facilitate the understanding of information representation in human brain.

### 1.1 Encoding models

In the previous literature, most encoding models were established based on specific computational rules. Neuroscientists believe that these computational rules may be the mathematical basis for the brain's response to specific visual stimuli.

For example, Kay *et al.* [1] used pyramid-shaped Gabor wavelet filters to build an encoding model. Based on this encoding model, the authors successfully identified the preferred natural images for the given human brain activities. Later, Kay *et al.* further proposed a two-stage cascade encoding model [6] based on the well-established local oriented filters, divisive normalization, compressive spatial summation and the variance-like nonlinearity. Recently, St-Yves and Naselaris [7] constructed a feature-weighted receptive field model based on the intermediate feature maps of a pretrained deep neural network (DNN), which can be used to predict the voxel response and study the shape of the receptive field of each voxel. Furthermore, Zeidman *et al.* [24] built a Bayesian population receptive field (pRF) model for interpretable brain encoding studies. In recent years, DNNs have made great success in computer vision, and researchers have begun to use DNNs to construct more complex brain encoding models [7, 21, 25]. In addition to encoding models for visual information, the researchers also studied how semantic information is expressed in the brain. For example, Huth *et al.* [26] established the mapping relationship between text semantic vectors and cerebral cortex activities, which presented us with a detailed semantic map of cerebral cortex.

### 1.2 Decoding models

Previous studies have demonstrated the feasibility of decoding the identity of binary contrast patterns [13-15], handwritten characters [16, 17], human facial images [18-20], natural picture/video stimuli [2, 21] and dreams [12, 22] from the corresponding brain activation patterns. For example, Miyawaki *et al.* [13] constructed a multiscale neural decoding model to reconstruct perceived binary contrast patterns from brain responses. Schoenmakers *et al.* [16] proposed a linear decoding model to reconstruct handwritten characters from brain responses. Güçlütürk *et al.* [20] proposed to combine probabilistic inference with adversarial training for reconstructions of perceived faces from brain responses. Horikawa *et al.* [2] showed that the hierarchical features of visual stimuli calculated by computer vision model could be predicted by utilizing the responses of multiple brain regions. This indicates that there is a close relationship between the hierarchical visual cortex and the complex visual features obtained by the computer vision model. Furthermore, Wen *et al.* [21] proposed a dynamic neural decoding method based on deep learning, which can reconstruct the dynamic visual scenes perceived by the human and predict its semantic labels. Horikawa *et al.* [22] even showed that brain activity could be used to predict the objects in humans' dreams.

Most of aforementioned decoding studies are based on multi-voxel pattern analysis (MVPA) method [8]. On the other hand, brain connectivity pattern is also a key feature of the brain state and can be used for brain decoding. Previous decoding studies [27-31] showed that brain connectivity information can been utilized as distinguishing features in procedures of decoding. For example, hiring brain connectivity information in brain decoding, Hossein et al. [30] successfully reconstructed two handwritten digits 6 and 9 from human brain activities. Manning *et al.* [31] proposed a probabilistic model for extracting dynamic functional connectivity patterns in brain activity. The proposed functional connectivity patterns can be used in brain decoding studies.

### 1.3 Hybrid encoding-decoding with bidirectional models

Although recent developments in brain encoding and decoding [6, 7, 13-22, 30, 32-34] have shown promising results, there are still many challenges in constructing an accurate decoding model to reconstruct the corresponding visual stimuli from fMRI data. From the Bayesian machine learning perspective, the encoding model can be acquired with a generative model that accounts for the measured brain activity. When it is combined with prior knowledge about the stimuli, a posterior probability distribution of the stimuli given a brain activity pattern, i.e., a predictive distribution for decoding, could be obtained. Therefore, the encoding and decoding models should not be taken as mutually exclusive. Effective unifying the encoding and decoding procedures may allow for accurate predictions and facilitate understanding of information representation in human brain [14, 35]. For example, Fujiwara *et al.* [14] proposed a "bidirectional" approach to visual image reconstruction, in which a set of latent variables were assumed to relate image pixels and fMRI voxels, and predictions for both encoding and decoding could be generated. They employed the Bayesian canonical correlation analysis (BCCA) framework, which computed multiple correspondences, via latent variables, between image pixels and fMRI voxels. Since the pixel weights for each latent variable can be thought to define an image basis, the training of the BCCA model using measured data leads to automatic extraction of image bases. Although it is premature to speculate on functional implications of the estimated image bases, this data-driven "bidirectional" approach could be extended to discover modular architecture of the brain in representing complex natural stimuli, behavior, and mental experience defined in high dimensional space.

## 2 Correspondence between DNNs and the human visual system

Deep learning [36, 37] is a large class of machine learning methods to extract hierarchical representations from input data. The architectures of DNN were first inspired by the structure and computational principles of biological nervous system [38]. Recently, DNN-based deep learning methods have achieved great success in image recognition, speech

recognition, natural language understanding and other aspects. In terms of architecture, the hierarchical layers of DNNs are very similar to that of the ventral visual system of human brain [36] (see Figure 2). In terms of function, existing researches on neural encoding and decoding based on deep learning have shown that the shallow representation of DNN is similar to the function of primary visual area, while the deep representation of DNN is similar to the back-end of ventral visual system [2, 25, 40, 41].

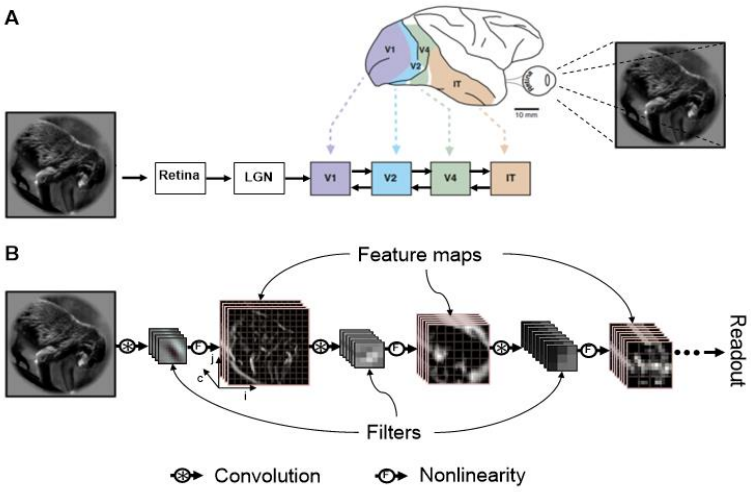


**Figure 2** The ventral visual system and the deep convolutional neural network (CNN). (**A**) The forward and backward projections between four Brodmann areas. (**B**) Illustration of a simple feedforward deep CNN, whose hierarchical structure is used to simulate the hierarchical representation of the ventral visual system. Adapted from [7, 45].

Humans can perceive the complex objects fast and accurately through the ventral visual stream, a system of interconnected brain regions that process increasingly complex features in hierarchical structures [39, 42, 43]. However, the automated discovery of early visual concepts from visual images without any supervised information is a major open challenge in machine perception research. On the one hand, we would like the representations extracted from the image perform well on real-world tasks. On the other hand, we would like to be able to interpret these representations, and they should be useful for tasks beyond those explicit in their initial design. Traditionally, it is difficult to adopt a pre-trained DNN model to learn such representations from visual images, because we don't know the semantic meaning of each dimensionality in the representation vector automatically extracted from the input image by that DNN model. Without disentangled representations, it is difficult to interpret these representations across different tasks. Fortunately, Higgins *et al.* [44] has shown that specially designed deep generative models are capable of learning disentangled representations

## 3 Brain decoding with deep generative models

It is a promising research direction to integrate deep learning methods into brain decoding research. Deep generative models such as variational autoencoders [46, 47] and generative adversarial networks [48] have achieved great success in the field of image generation. Recently, more and more attention has been paid to the research of visual image reconstruction using deep generative models [20, 32-34, 49, 50].

### 3.1 VAE based methods

Variational autoencoders (VAE), originally presented in [46, 47], is a probabilistic extension of auto-encoder model. It has a bottom-up encoding network and a top-down decoding network. These two networks are jointly trained to maximize the lower bound of the data likelihood, thereby reformulating the auto-encoder model as a variational inference problem. Recent works have demonstrated that the VAE based models are capable of learning disentangled representations that correspond to distinct factors of variation in the input data [44, 51, 52]. That is very important for brain encoding and decoding tasks, since some visual concepts learned by VAE based models are also perceived by the human brain. Inspired by this, researchers explored the VAE based models for image reconstruction from brain activities [32, 33].

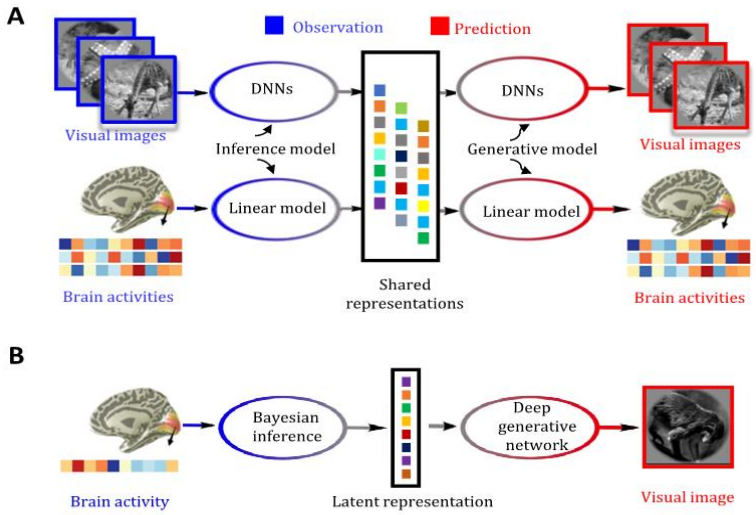


**Figure 3** Illustration of the deep generative multi-view framework for neural decoding. (**A**) Model training. View-specific generative models are used for data generation. Specifically, DNN is adopted to model visual images, while linear regression model is used to model brain activities. (**B**) Image reconstruction. The brain activities independent of those used for the training are decoded to visual images. Cited from [32].

For example, Du *et al.* [32] proposed a deep generative multi-view model (DGMM) for reconstructing the perceived images from brain fMRI activities (see Figure 3). DGMM can be viewed as a nonlinear extension of the linear BCCA. Under the DGMM framework, the encoding and decoding procedures are simultaneously formulated by two distinct generative models:

$$p_\theta(\mathbf{X}\,|\,\mathbf{Z}) = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{x}_i \,|\, \mu_{\mathbf{x}}(\mathbf{z}_i), \mathrm{diag}(\sigma^2_{\mathbf{x}}(\mathbf{z}_i))\right), \tag{1}$$

$$p(\mathbf{Y}\,|\,\mathbf{Z}) = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{y}_i \,|\, \mathbf{B}^\top \mathbf{z}_i,\ \Psi\right), \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{D_x \times N}$ denotes the visual images, $\mathbf{Y} \in \mathbb{R}^{D_y \times N}$ denotes the evoked fMRI activities, $\mathbf{Z} \in \mathbb{R}^{D_z \times N}$ denotes the shared latent variables between the visual images and the evoked fMRI activities. The training set consists of $N$ paired samples, which can be denoted by $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x}_i \in \mathbb{R}^{D_x}$ and $\mathbf{y}_i \in \mathbb{R}^{D_y}$ for $i = 1, \ldots, N$. Specifically, the DGMM uses a deep neural network based generative process to model the distribution of visual images, while uses a sparse linear generative process to model the distribution of brain response data. On the one hand, the deep neural network used here can effectively capture the hierarchical features of the visual image, which is similar to the structure of the ventral visual system of the human brain [2, 25, 40, 41]. On the other hand, the sparse linear generative model used here not only conforms to the sparse expression principle of human brain, but also avoids over-fitting of brain response data [53]. Note that these two generative processes share the same latent variables. Therefore, in test phase, it allows us to infer the corresponding visual image from the brain response through the shared latent variables. Actually, the DGMM framework can capture the "bidirectional" mapping relationships between the visual images and the corresponding fMRI activities. Thanks to its auto-encoding variational Bayesian architecture, the DGMM can be optimized efficiently by means of mean-field variational inference, which is similar to the classical VAE solution. Compared with the non-probabilistic deep multi-view learning methods, the DGMM's Bayesian framework makes it naturally more flexible and adaptive.

### 3.2 GAN based methods

Generative adversarial network (GAN) is first proposed in [48]. The basic GAN is an unsupervised model that generates images from noise vector. The idea of adversarial training comes from the game theory, where two competitors compete with each other to make progress together. The typical configuration of GAN is composed with a generator and a discriminator. The task of the generator is to synthesize images from noise to deceive discriminator into believing that the synthesized images are real-world scenes. Meanwhile, the discriminator attempts to distinguish between the synthesized data and real data. When the Nash equilibrium is reached, the generator learns the distribution of the real-world images, and the discriminator is sensitive to capture the difference between real and fake data. GAN has been widely used in various applications, e.g., image generation [54], image-to-image translation [55] and text-to-image synthesis [56, 57].

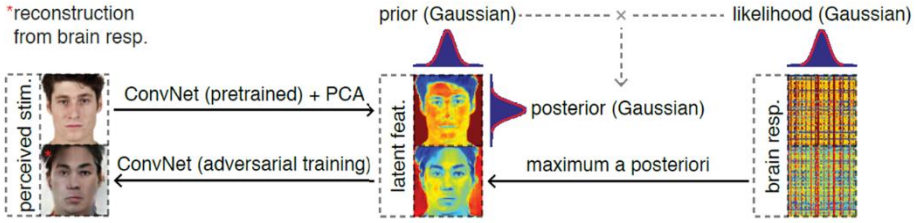


**Figure 4** Illustration of the deep adversarial neural decoding. By combining probabilistic inference with adversarial learning, it can clearly reconstruct the corresponding face image from the brain activity. Cited from [20].

Unlike VAE, GAN is a likelihood-free model, i.e., it does not make any prior assumptions on the data distribution, and the data distribution is totally learned via adversarial training. This is a favorable characteristic in neural encoding and decoding tasks. GAN often requires exact semantic information flow in its generator and discriminator. However, the useful semantic information in the blood oxygen level-dependent (BOLD) signal is merged deep in noise, which is a great challenge for the model training. A recent brain decoding research [20] proposes to combine probabilistic inference with adversarial training for reconstructions of perceived faces from brain activations (see Figure 4). Assume that $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ is the visual image, $\mathbf{z} \in \mathbb{R}^{p}$ is its latent features, $\mathbf{y} \in \mathbb{R}^{q}$ is the corresponding brain response, and $\phi : \mathbb{R}^{h \times w \times c} \to \mathbb{R}^{p}$ is a latent feature model such that $\mathbf{z} = \phi(\mathbf{x})$ and $\mathbf{x} = \phi^{-1}(\mathbf{z})$. Then the perceived visual images can be reconstructed from brain responses by the following equation:

$$\mathbf{x} = \phi^{-1}\left( \underset{\mathbf{z}}{\operatorname{argmax}}\, p(\mathbf{z} \mid \mathbf{y}) \right), \tag{3}$$

where $p(\mathbf{z} \mid \mathbf{y})$ is the posterior distribution of the latent variables. Equation (3) can be reformulated through Bayes' theorem:

$$\mathbf{x} = \phi^{-1}\left( \underset{\mathbf{z}}{\operatorname{argmax}} \left[ p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}) \right] \right), \tag{4}$$

where $p(\mathbf{y} \mid \mathbf{z})$ is the likelihood function, and $\mathrm{p}(\mathbf{z})$ is the prior distribution of the latent variables. Intuitively, the authors first decode the observed brain responses to the latent features with maximum a posteriori estimation. Then, they generate the perceived images according to the decoded latent features with adversarial learning. This two-step brain decoding method can accurately generate the reconstructions of perceived faces from the brain responses. More recently, researchers try to reconstruct the natural images from the measured fMRI signals [34, 49, 50] by utilizing GANs which are pre-trained on the large-scale image datasets.

## 4 Improving brain encoding and decoding with dual learning

Data-driven brain encoding and decoding methods often require acquisition of a large number of paired (stimulus-response) data instances to train a model customized to individual subject. In many encoding and decoding studies, however, one can gather at most a few thousand noisy paired data instances from one subject. To improve the generalization ability of the encoding and decoding models, one therefore needs to make good use of large-scale unpaired data instances (e.g., visual images).

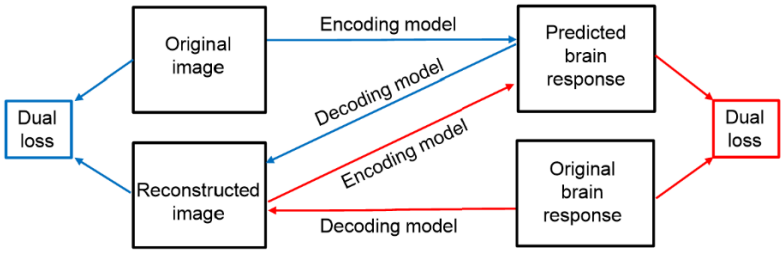


**Figure 5** Improving brain encoding and decoding with dual learning. The dual loss measured over unpaired data (either visual images or brain responses) would generate informative feedback to train the bidirectional mapping model. Under this dual learning framework, we can leverage the large-scale unpaired data to improve models' generalization ability.

Inspired by recently proposed dual learning for machine translation [58, 59], we can train the encoding and decoding models simultaneously by minimizing the reconstruction loss resulting from the bidirectional mapping model. The

encoding and decoding models represent a primal-dual pair and form a closed loop, allowing the application of dual learning (see Figure 5). Specifically, the reconstruction loss measured over unpaired data (e.g., visual images) would generate informative feedback to train the bidirectional mapping model. Under this dual learning framework, we can leverage the large-scale unpaired visual images to improve the generalization ability of the encoding and the decoding models. Actually, dual learning is a general framework for learning the bidirectional mappings from one data domain X to another data domain Y [60, 61]. For $X \rightarrow Y$, the goal is to learn a encoder mapping E such that the distribution E(X) is indistinguishable from the distribution Y using an adversarial loss. Similarly, for $Y \rightarrow X$, the goal is to learn a decoder mapping D such that the distribution D(Y) is indistinguishable from the distribution X using another adversarial loss. In particular, on the paired data, one can combine these two adversarial losses and the cycle consistency losses (dual losses) to push $D(E(X)) \approx X$ and $E(D(Y)) \approx Y$.

## 5 Conclusions

In conclusion, brain encoding and decoding are central to the field of computational neuroscience and has the potential to create better brain-machine interfaces. The architecture and computational rule of deep neural networks share some similarity with the human visual streams. The use of deep generative models (e.g., VAEs and GANs) in brain encoding and decoding studies holds promise to provide deeper insight into relationships between human visual experience and the evoked neural activity. By leveraging the large-scale unpaired data, dual learning is expected to play an important role in developing neural encoding and decoding models.


## Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 91520202), CAS Scientific Equipment Development Project (YJKYYQ20170050), Beijing Municipal Science & Technology Commission (Z181100008918010), Youth Innovation Promotion Association CAS and Strategic Priority Research Program of CAS.


## Compliance with ethics guidelines

Changde Du, Jinpeng Li, Lijie Huang and Huiguang He declare that they have no conflict of interest or financial conflicts to disclose.

## References


[1] K. N. Kay, T. Naselaris, R. J. Prenger, J. L. Gallant, Identifying natural images from human brain activity, Nature, 2008, 452 (7185) : 352–355.

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

[3] T. Naselaris, K. N. Kay, S. Nishimoto, J. L. Gallant, Encoding and decoding in fMRI, NeuroImage, 2011, 56 (2): 400–410.

[4] M. Chen, J. Han, X. Hu, X. Jiang, L. Guo, and T. Liu. Survey of encoding and decoding of visual stimulus via fMRI: an image analysis perspective. Brain imaging and behavior, 2014, 8(1): 7–23.

[5] M. A. van Gerven. A primer on encoding models in sensory neuroscience. Journal of Mathematical Psychology, 2017, 76: 172–183.

[6] K. N. Kay, J. Winawer, A. Rokem, A. Mezer, and B. A. Wandell. A two-stage cascade model of bold responses in human visual cortex. PLoS Comput Biol, 2013, 9(5): e1003079.

[7] G. St-Yves and T. Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. NeuroImage, 2018, 180 (Part A): 188-202

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

[9] J. Haynes and G. Rees. Decoding mental states from brain activity in humans. Nature Reviews Neuroscience, 2006, 7(7):523–534.

[10] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant. Bayesian reconstruction of natural images from human brain activity. Neuron, 2009, 63(6):902–915.

[11] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fMRI. Neuroimage, 2011, 56(2):400–410.

[12] T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani. Neural decoding of visual imagery during sleep. Science, 2013, 340(6132):639–642.

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

[14] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani. Modular encoding and decoding models derived from Bayesian canonical correlation analysis. Neural computation, 2013, 25(4):979–1005.

[15] S. Yu, N. Zheng, Y. Ma, H. Wu, and B. Chen. A novel brain decoding method: a correlation network framework for revealing brain connections. arXiv preprint arXiv:1712.01668, 2017.

[16] S. Schoenmakers, M. Barth, T. Heskes, and M. van Gerven. Linear reconstruction of perceived images from human brain activity. NeuroImage, 2013, 83:951–961.

[17] S. Schoenmakers, U. Güc¸lü, M. Van Gerven, and T. Heskes. Gaussian mixture models and semantic gating improve reconstructions from human brain activity. Frontiers in computational neuroscience, 2015, 8: 173.

[18] A. S. Cowen, M. M. Chun, and B. A. Kuhl. Neural portraits of perception: reconstructing face images from evoked brain activity. NeuroImage, 2014, 94:12–22.

[19] H. Lee and B. A. Kuhl. Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. The Journal of Neuroscience, 2016, 36(22):6069–6082.

[20] Y. Güçlütürk, U. Güçlü, K. Seeliger, S. Bosch, R. van Lier, and M. A. van Gerven. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. Advances in Neural Information Processing Systems (NIPS), 2017: 4249-4260

[21] H. Wen, J. Shi, Y. Zhang, K. Lu, J. Cao, and Z. Liu. Neural encoding and decoding with deep learning for dynamic natural vision. Cerebral cortex, 2017, 1–25,

[22] T. Horikawa and Y. Kamitani. Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. Frontiers in computational neuroscience, 2017, 11:4.

[23] T. Naselaris, C. A. Olman, D. E. Stansbury, K. Ugurbil, and J. L. Gallant. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. Neuroimage, 2015, 105:215–228

[24] P. Zeidman, E.H. Silson, D.S. Schwarzkopf, C.I. Baker, W. Penny. Bayesian population receptive field modelling. Neuroimage, 2018, 180 (Part A):173–187.

[25] U. Güçlü and M. A. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience, 2015, 35(27):10005–10014.

[26] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, J. L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 2016, 532: 453–458.

[27] W. Shirer, S. Ryali, E. Rykhlevskaia, V. Menon, and M. D. Greicius. Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cerebral cortex, 2012, 22(1):158–165.

[28] F. Mokhtari and G. Hossein-Zadeh. Decoding brain states using backward edge elimination and graph kernels in fMRI connectivity networks. Journal of neuroscience methods, 2013, 212(2):259–268.

[29] E. Yargholi and G. A. Zadeh. Brain decoding-classification of hand written digits from fMRI data employing Bayesian networks. Frontiers in human neuroscience, 2016, 10:351.

[30] G. Hossein-Zadeh et al. Reconstruction of digit images from human brain fMRI activity through connectivity informed Bayesian networks. Journal of neuroscience methods, 2016, 257:159–167.

[31] J.R. Manning, X. Zhu, T.L. Willke, R. Ranganath, K. Stachenfeld, U. Hasson, D.M. Blei, K.A. Norman. A probabilistic approach to discovering dynamic full-brain functional connectivity patterns. Neuroimage, 2018, 180 (Part A): 243–252.

[32] C. Du, C. Du, and H. He. Sharing deep generative representation for perceived image reconstruction from human brain activity. In International Joint Conference on Neural Networks (IJCNN), 2017: 1049-1056.

[33] K. Han, H. Wen, J. Shi, K. Lu, Y. Zhang, and Z. Liu. Variational autoencoder: An unsupervised model for modeling and decoding fMRI activity in visual cortex. bioRxiv, 2017, page 214247.

[34] K. Seeliger, U. Güc¸lü, L. Ambrogioni, Y. Güc¸lütürk, and M. A. van Gerven. Generative adversarial networks for reconstructing natural images from brain

activity. NeuroImage, 2018, 181: 775-785.

[35] P. Kuo, Y. Chen, L. Chen, and J. Hsieh. Decoding and encoding of visual patterns using magnetoencephalographic data represented in manifolds. NeuroImage, 2014, 102:435–450

[36] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. Nature, 2015, 521(7553):436–444.

[37] J. Schmidhuber. Deep learning in neural networks: An overview. Neural networks, 2015, 61:85–117.

[38] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 1943, 5(4):115–133.

[39] J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? Neuron, 2012, 73(3):415–434.

[40] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific reports, 2016, 6.

[41] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage, 2017, 152:184–194.

[42] J. J. DiCarlo and D. D. Cox. Untangling invariant object recognition. Trends in cognitive sciences, 2007, 11(8):333–341.

[43] J. Li, Z. Zhang, and H. He. Visual information processing mechanism revealed by fMRI data. In International Conference on Brain and Health Informatics, 2016, pages 85–93.

[44] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, A. Lerchner, Early visual concept learning with unsupervised deep learning, arXiv preprint arXiv:1606.05579.

[45] D. D. Cox and T. Dean. Neural networks and neuroscience-inspired computer vision. Current Biology, 2014, 24(18): 921–929.

[46] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In International Conference on Learning Representations (ICLR), 2014.

[47] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Advances in Neural Information Processing Systems (NIPS), 2014, 1278–1286.

[48] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS), 2014, pages 2672–2680.

[49] G. St-Yves and T. Naselaris. Generative adversarial networks conditioned on brain activity reconstruct seen images. bioRxiv, 2018, page 304774.

[50] G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani. End-to-end deep image reconstruction from human brain activity. bioRxiv, 2018, page 272518.

[51] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In Advances in Neural Information Processing Systems (NIPS), 2015, pages 2539–2547.

[52] S. A. Eslami, N. Heess, T. Weber, Y. Tassa, D. Szepesvari, G. E. Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In Advances in Neural Information Processing Systems (NIPS), 2016, pages 3225–3233.

[53] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences, 2006, 10(9):424–430.

[54] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004, 2016

[55] M. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In Advances in Neural Information Processing Systems (NIPS), 2017, pages 700–708.

[56] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396, 2016.

[57] S. Hong, D. Yang, J. Choi, and H. Lee. Inferring semantic layout for hierarchical text-to-image synthesis. arXiv preprint arXiv:1801.05091, 2018.

[58] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W. Ma. Dual learning for machine translation. In Advances in Neural Information Processing Systems (NIPS), 2016, pages 820–828.

[59] Y. Xia, T. Qin, W. Chen, J. Bian, N. Yu, and T. Liu. Dual supervised learning. International Conference on Machine Learning (ICML), 2017: 3789-3798

[60] Y. Xia, X. Tan, F. Tian, et al. Model-level dual learning, International Conference on Machine Learning (ICML), 2018: 5379-5388.

[61] J. Zhu, T. Park, P. Isola, A. A. Efros: Unpaired image-to-Image translation using cycle-consistent adversarial networks. International Conference on Computer Vision (ICCV), 2017: 2242-2251.