Rethinking Visual Reconstruction: Experience-Based Content Completion Guided by Visual Cues

Jiaxuan Chen [1 2 *] Yu Qi [1 3 *] Gang Pan [1 2]

Abstract

Decoding seen images from brain activities has been an absorbing field. However, the reconstructed images still suffer from low quality with existing studies. This can be because our visual system is not like a camera that "remembers" every pixel. Instead, only part of the information can be perceived with our selective attention, and the brain "guesses" the rest to form what we think we see. Most existing approaches ignored the brain completion mechanism. In this work, we propose to reconstruct seen images with both the visual perception and the brain completion process, and design a simple, yet effective visual decoding framework to achieve this goal. Specifically, we first construct a shared discrete representation space for both brain signals and images. Then, a novel self-supervised token-to-token inpainting network is designed to implement visual content completion by building context and prior knowledge about the visual objects from the discrete latent space. Our approach improved the quality of visual reconstruction significantly and achieved state-of-the-art.

1. Introduction

Seeking the relationship between brain activities and the corresponding visual stimulus is an interesting topic in neural decoding, which not only contributes to the development of intelligence paradigms (Wu et al., 2013; Yu et al., 2016; Wang et al., 2015) but also provides vital application values, e.g., neural-prostheses (Qian et al., 2020). Existing studies have shown that using functional Magnetic Resonance Imaging (fMRI), which is a non-invasive technique to measure brain activities, can effectively reconstruct perceived information with deep neural networks (Shen et al., 2019a; Beliy et al., 2019; Fang et al., 2020; Mozafari et al., 2020; Ren et al., 2021; Gaziv et al., 2022; Ozcelik et al., 2022). For the existing visual reconstruction approaches, one critical problem lies in that they usually fail to recover the intricate color and texture in natural scenes, or the reconstructed results are often unfaithful to the real images, especially for generative adversarial networks (GANs)-based approaches.

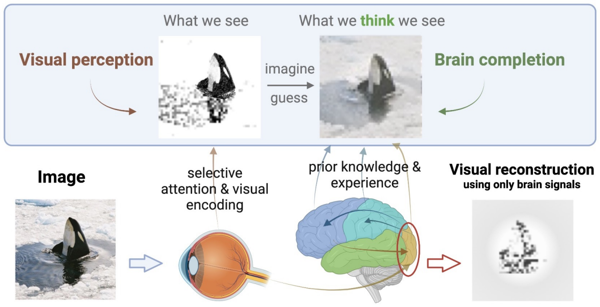

Figure 1. What we see is a combination of both the visual perceptions (which can be limited, focusing on certain parts mostly) and the brain completion process (according to our knowledge and experience). Therefore, using brain signals alone is not sufficient to reconstruct the seeing image comprehensively.

To cope with the existing dilemma in visual decoding, we need to ask: *is the use of fMRI recording alone sufficient to reconstruct the perceived image with details?* One limitation may lie in the capacity for preserving brain activities of fMRI signals, while another factor is whether our visual system conveys the full visual information to the brain. Unfortunately, our visual system is not like a camera that "remembers" every pixel of seen images, and it is usually difficult for us to describe the detail of a seen scene. In fact, the visual information provided to the retina is limited due to the selective visual attention mechanism, such that only a certain part of the information that interests us is perceived by the eyes (Desimone et al., 1995). Furthermore, the number of ganglion cells is far fewer than the photoreceptor cell, which causes the visual stimulus transmitted

[*] Equal contribution [1] State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China. [2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China. [3] MOE Frontier Science Center for Brain Science and Brain-Machine Integration, Zhejiang University, Hangzhou, China. Correspondence to: Yu Qi <qiyu@zju.edu.cn>, Gang Pan <gpan@zju.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

to the central nervous system will be further compressed (Gazzaniga, 2009).

But why do we *think* that we perceive everything we see? From the view of neuroscience, our visual perception is an active and creative process and has constructive nature (Kandel et al., 2000), namely, the brain will conjecture the scene presented to the eyes by the incoming stream of visual signals and the past experience (e.g., learned regularities of the world, and appropriate frame of reference (Hinton & Lang, 1985)), as shown in Fig. 1. In other words, our visual perception worlds rely not only on lossy visual information from the retina, but also on cognitive function driven by the experience. Thus, only using the brain activities maybe not be sufficient to reconstruct seen images in detail, especially for color, texture, and background information.

To fulfill the aforementioned objectives, we propose to reconstruct seen images with both the visual perception and the brain completion process, and design a novel fMRI-to-image reconstruction framework (VQ-fMRI). Firstly, we learn discrete visual representations and constituent contexts of images in a self-supervised manner, which is regarded as the process of building a visual experience. Then, we seek common visual cues between fMRI and images under a set of shared prototype vectors, such that brain signals (visual perception) and experiential images (knowledge) can be matched on a shared representation space. Based on this, given visual cues from brain signals, we can infer the uncertain or missing content by an experience-based completion model with the image set, guided by the known information. Further, we design a hierarchical architecture to improve the quality of reconstructed images via alternating compression and super-resolution steps.

The core part of our approach is the content completion process, requiring the prediction of the visual content to be harmonized with the given cues decoded from fMRI, which is achieved by a novel token-based inpainting model. Note that token represents the index of a specific visual vector in the discrete representation space. Before token inpainting, the proposed VQ-fMRI provides a mechanism to “transform” multimodal data of fMRI and images into canonical visual tokens. Therefore, by capturing and understanding the context interrelations of discrete token sequences, where a sequence of visual tokens can well represent the intrinsic structure of images (Van Den Oord et al., 2017; Esser et al., 2021), it is possible to expect the recalibration region of visual content to satisfy coordination with the known discrete cues and is semantically plausible.

Our contributions can be summarized as three-fold:

- We propose a novel Vector-Quantization fMRI decoding model (VQ-fMRI), which formulates visual reconstruction as experience-based context completion guided by visual cues from brain activities, to investigate the feasibility of simulating a brain-like visual perception mechanism.
- We propose a cross-modal inpainting self-supervised framework, providing a foundation for fulfilling decoding verification and deviation correction. This model allows us to avoid focusing or spending capacity on decoding imperceptible local details.
- Compared with previous leading methods, the images reconstructed by our approach are more faithful to the stimulus images with better preserved low-level color textures, and high-level semantic information.

2. Related Work

Linear Model. Early approaches primarily focus on estimating a linear mapping between fMRI voxels and handcrafted image features (Miyawaki et al., 2008; Schoenmakers et al., 2013). The handcrafted descriptor is designed to mimics the brain activity in the visual cortex, and then the decoding target can be achieved via predicting the responses of each voxel from the handcrafted features, or mapping voxel responses to image features (both in linear). Although reconstructing with a simple linear regression model has gained satisfactory performance for low-level detail stimuli, it struggles to reconstruct complex natural images, and the performance still lags behind the advanced deep learning-based alternatives. The reason behind this could be that the linear hypothesis is not enough to correctly express the encoding and decoding rules in the human visual system.

Learning-based Decoding. Recently, solving natural image reconstruction with deep neural networks (DNNs) has received a lot of interest. Motivated by the fact that the hierarchical feature of CNNs correlates with brain visual activity, Shen *et al.* (Shen et al., 2019b) exploited a pre-trained VGG-19 model (Simonyan & Zisserman, 2014) to yield neural representations, and then optimized the input image for minimizing the difference between the DNN representations and the decoded fMRI features. A concurrent work (Shen et al., 2019a) also used pre-trained visual representation proxy, but designed an end-to-end reconstruction model. Instance-conditioned GAN (IC-GAN) (Casanova et al., 2021), a novel GAN technique, which is recently introduced to guide the training of a ridge regression model (Ozcelik et al., 2022), where the regression model aims to decode latent variables of a pre-trained IC-GAN from the fMRI patterns. Since the publication of latent diffusion model (LDM) (Rombach et al., 2022), many diffusion-based visual reconstruction methods have been emerged (Takagi & Nishimoto, 2023; Chen et al., 2023). A stronger generative model can improve reconstructing performance, but the key problem lies in how to guarantee that the generated images

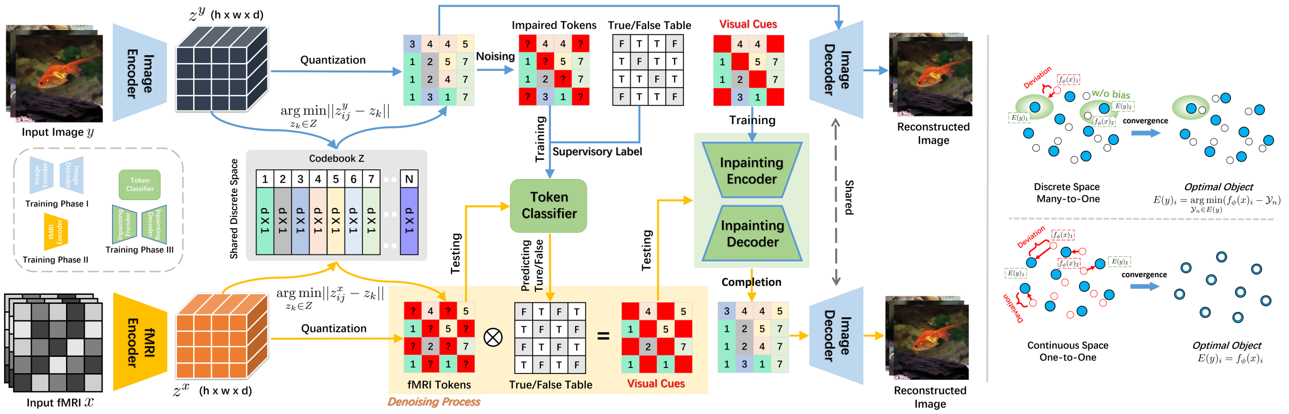

Figure 2. The proposed VQ-fMRI framework. In the training phase, we first leverage a VQ-VAE to learn image feature constituents (codebook). Next, training a fMRI encoder for mapping voxel vectors to the discrete tokens, which specifies the entries in the codebook, guided by the corresponding image tokens. Meanwhile, a large number of image tokens are created from ImageNet, and a subset of tokens is selected as unknown content by random replacement. These corrupted tokens are fed into a classifier to encode the confidences, and then the unknown content is inferred by an inpainting model. Right: the difference between discrete and continuous feature proxy, where $f_\phi(\cdot)$ is a mapping with learnable parameters, $E(\cdot)$ and $\mathcal{Y}_n$ denotes the pre-trained network and learned image features, respectively.

contain the low-level features of the visual stimuli. On the other hand, Beliy *et al.* (Beliy et al., 2019) first proposed a self-supervision visual decoding framework. The main principle is introducing a image-to-fMRI encoder, and a fMRI-to-image decoder network, and then concatenating back to back into two symmetric architectures: encoder-decoder, and decoder-encoder. This design allows training on larger unlabeled fMRI and image datasets, but the separate training strategy is prone to catastrophic forgetting problems. An improved version (Gaziv et al., 2022), providing new reconstruction and classification capabilities, is also developed recently. Additionally to the approaches mentioned above, several relevant studies also include (Du et al., 2022; St-Yves & Naselaris, 2018; Qiao et al., 2020; Ren et al., 2021; Mozafari et al., 2020).

3. Method

Below, let us introduce our VQ-fMRI from three main aspects: modeling of visual cues, token inpainting module, as illustrated in the Fig. 2, and hierarchical super-resolution architecture (see Fig. 4). In the remainder of this section, the problem statement is first discussed. Subsequently, we elaborate the concrete implementations of VQ-fMRI.

Problem Statement. Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ represents the fMRI-image dataset, where x_i denotes fMRI recording, and y_i is the corresponding visual image. To meet visual decoding, a simple idea is to seek a mapping $f_\theta : x_i \to y_i$ by minimizing $\mathbb{E}_{x_i, y_i \sim \mathcal{D}} ||y_i - f_\theta(x_i)||_2^2$. Nevertheless, since $\mathcal{D}$ is relatively limited, reconstruction directly from fMRI data is considered infeasible (Shen et al., 2019a). In contrast, the more promising practice is to learn shared neural representations for both fMRI and image, and the fMRI latent representation $\mathcal{X}_i$ is usually guided via a a pre-trained teacher network, i.e., learning a mapping $f_\phi : x_i \to \mathcal{X}_i, s.t.\ \mathcal{X}_i = \mathcal{Y}_i$, where $\mathcal{Y}_i$ denotes the intermediate features of y_i from a pre-trained network (Shen et al., 2019b; Du et al., 2022; Fang et al., 2020; Du et al., 2018; Ren et al., 2021; Ozcelik et al., 2022). Then, $\mathcal{X}_i$ is fed to a decoding model to generate visual images. To generate more realistic images, advanced generative adversarial nets (Goodfellow et al., 2014; Isola et al., 2017; Brock et al., 2018; Casanova et al., 2021) are widely adopted. However, there are also the following limitations: i) it is challenging to accurately align the potential representations of (x_i, y_i) in continuous space, and ii) the gap between them leads to reconstruction results unfaithful to the raw stimulus images. This can be solved by increasing the training samples, but the cost of collecting labeled samples is enormous. Therefore, we seek to propose a new self-supervised approach, inspired by cognitive neural science, to revisit visual neural decoding. For clarity, we only use the x and y to denote fMRI and image in the following (unless otherwise noted).

3.1. Modelling of Visual Cues

As aforementioned, the optimization on continuous space is prone to cumulative errors due to the mistakes in the earlier representation learning, and how such prediction bias affects subsequent reconstruction task is difficult to quantify. In order to establish reliable visual cues, therefore, we recommend to express the constituents of an image in the form of discrete prototype vectors. Unlike the continuous feature space that requires strong alignment in all dimensions between representations, discretization allows us to relax the constraints of one-to-one coordination in shared embedding space, thus relieving the potential disturbances from the noise present in the fMRI data (see Fig. 2 right).

Learning Codebook. To learn a discrete embedding space, a VQ-VAE (Van Den Oord et al., 2017), comprising of an encoder $\mathbf{E}$, a decoder $\mathbf{D}$, and a discrete codebook $Z = \{z_k\}_{k=1}^{K} \in \mathbb{R}^{k\times d}$, is employed. Note that K denotes the number of prototypes, and d represents the dimension of codes. More formally, given an image y, the output of encoder $z^y = \mathbf{E}(y) \in \mathbb{R}^{h\times w\times d}$ is passed through a element-wise quantization $\mathbf{VQ}(\cdot)$ producing the spatial collection of image tokens $\mathcal{Z}^y \in \mathbb{R}^{h\times w\times d}$:

$$\mathcal{Z}^y = \mathbf{VQ}(z^y; Z) := \Big(\underset{z_k\in Z}{\arg\min}||z_{ij}^y - z_k||\Big). \tag{1}$$

Intuitively, $\mathbf{VQ}(\cdot)$ maps each learned spatial code z_{ij}^y to the nearest prototype vector in the Z, where $h \times w$ denotes the sequence length and is usually much smaller than the original image. Then, decoder $\mathbf{D}$ can be used to recover the observation y faithfully, i.e., $y^* = \mathbf{D}(\mathbf{VQ}(\mathbf{E}(y)))$, by optimizing the following objective in an end-to-end manner:

$$\begin{aligned}\mathcal{L}_{vq} = ||y^* - y||_2^2 &+ ||\text{sg}[\mathbf{E}(y)] - \mathcal{Z}^y||_2^2 \\ &+ \beta||\text{sg}[\mathcal{Z}^y] - \mathbf{E}(y)||_2^2.\end{aligned} \tag{2}$$

In Eq. 2, the first term $||y^* - y||_2^2$ represents reconstruction loss, and $\beta||\text{sg}[\mathcal{Z}^y] - \mathbf{E}(y)||_2^2$ is the commitment loss, where $\text{sg}[\cdot]$ refers to a stop-gradient operation.

Discrete Visual Cues. Once a well-trained VQ-VAE has been acquired, our goal (i.e., visual cues modeling) is transformed into a K-way classification problem, where the discrete visual parts in the codebook Z is the potential candidates. We propose a lightweight convolutional model to implement such a "domain migration". In order of computation, the process of fMRI embedding is split into three parts. Firstly, a multilayer perceptron (MLP) takes the input x, and through 2 hidden layers outputs a feature map z_*^x (constrained to be the same size as the z^y). Next, z_*^x is fed into a U-Net with skip connections, which can effectively preserve and fuse both low and high-level abstract features, thereby more information can be passed from the fMRI voxel space to the discrete embedding space. In the end, the output z^x of U-Net is quantized based on its distance to the codebook entries:

$$\mathcal{Z}^x = \mathbf{VQ}\Big(\mathbf{UNet}(\text{MLP}(x)); Z\Big) \in \mathbb{R}^{h\times w\times d}. \tag{3}$$

The above discrete embedding learning can be viewed as a hard clustering operation, which relaxes the constraint that latent variables must be equal in all dimensions. Now, the identical codebook entries in the same positions can be used to build visual cues:

$$\underset{z_k\in Z}{\arg\min}||z_{ij}^x - z_k|| = \underset{z_k\in Z}{\arg\min}||z_{ij}^y - z_k||. \tag{4}$$

Loss Function. Learning-based fMRI-to-image methods commonly combine a mean square error (MSE) loss to maximize the similarity between shared latent representations. However, minimizing MSE may suffer from "regression-to-the-mean" issue. To mitigate this, we propose a simple VQ-MSE loss, which can be formulated as:

$$\begin{aligned}&\mathcal{L}_{vm}(z^x, z^y) = \sum_{i=1}^{h}\sum_{j=1}^{w} I_{ij}(z^x, z^y)\Big|\Big|z_{ij}^x - [\mathbf{VQ}(z^y)]_{ij}\Big|\Big|_2^2, \\ &I_{ij}(z^x, z^y) = \begin{cases} 0, & \text{if } \underset{z_k\in Z}{\arg\min}||z_{ij}^x - z_k|| = [\mathbf{VQ}(z^y)]_{ij} \\ 1, & \text{otherwise} \end{cases}\end{aligned} \tag{5}$$

The intuition behind is our training objective penalizes only spatial codes that are mapped to incorrect nearest neighbor prototypes, while small perturbations occurring around the correct prototype do not alter the loss value, as shown in Fig. 2 right.

3.2. Token-to-Token Inpainting Based on Known Cues

Given an image with random masked patches, the human brain can always imagine the occluded part from the visible region (prior). In the CV community, this task has been investigated in different contexts (e.g., colorization, and uncropping), and achieved satisfactory performance (Liu et al., 2020; Esser et al., 2021; Lugmayr et al., 2022; Saharia et al., 2022). These techniques, however, are ill-suited for applying directly on the visual reconstruction tasks due to the heterogeneity (e.g., distribution) between image and fMRI. To the best of our knowledge, a similar mechanism has not been adequately explored in the field of neural decoding. The proposed token-to-token inpainting framework (Fig. 2 middle) is expected to fill this gap.

With the visual cues definition in Eq. 4, any fMRI embedding result can be considered a matrix $\mathcal{M}^x \in \mathbb{R}^{h\times w}$ (only contains 0 or 1), where 1 indicates the correctly decoded token, and 0 is false prediction. Subsequently, the known cues can be extracted by $\mathcal{Z}_{\text{T}}^x = \mathcal{M}^x \odot \mathcal{Z}^x$ ($\odot$ is Hadamard product). Under the circumstances, our inpainting task boils down to correcting the mismatches $\mathcal{Z}_{\text{F}}^x = (1 - \mathcal{M}^x) \odot \mathcal{Z}^x$ by conditioning on $\mathcal{Z}_{\text{T}}^x$. How to get a large number of visual cues? Gathering from the fMRI training set is obviously not enough. A reasonable practice is to obtain such prior information from the images, which is also the superiority of our token-based model.

To repair fMRI embedding tokens that lead to visual disharmony, our learning strategy is straightforward: we use the pre-trained VQ-VAE to produce quantized latent variables of images, and sample random (following a uniform distribution) prototype vectors from the codebook to replace these encoded latent variables. The proportion of replacement is comparable to the correct rate of prediction from fMRI data. On the one hand, the elements of $\mathcal{M}^x$ cannot be determined during the forward propagation, so we can only draw on random replacement instead of direct masking. The replaced location records $\mathcal{M}^y \in \mathbb{R}^{h\times w}$, on the other hand, are available in the training phase. Therefore, our inpainting network consists of two modules: the former aims to differentiate between real tokens and random tokens (outputting the corresponding confidence score $\mathcal{M}^y_* \in \mathbb{R}^{h\times w}$), and the latter recalibrates the missing latent variables from $\mathcal{M}^y_* \odot \mathcal{Z}^y_{\mathrm{R}}$ (where $\mathcal{Z}^y_{\mathrm{R}}$ denotes the replaced embedding), as shown in Fig. 2 middle. Note that $\mathcal{M}^y_* \odot \mathcal{Z}^y_{\mathrm{R}}$ is equivalent to the masking operation. Our overall training objective can be expressed as:

$$\mathcal{L}_{fix} = \mathcal{L}_{bce}(\mathcal{M}^y_*, \mathcal{M}^y) + \lambda \mathcal{L}_{vm}(\mathcal{Z}^y_{\mathrm{R}}, \mathcal{Z}^y), \tag{6}$$

where $\mathcal{L}_{bce}$ denotes binary cross-entropy loss, and $\lambda = 2$ is tradeoff parameter. In principle, our framework involves forward noising process $p(\mathcal{Z}^y_{\mathrm{R}}|\mathcal{Z}^y)$, and reverse inference process $p(\mathcal{Z}^y|\mathcal{Z}^y_{\mathrm{R}})$. This problem is well suited to be modeled via a reverse Markov chain, i.e., diffusion model (Ho et al., 2020) for iteratively recovering information from noise. However, we leave it to future work, and the main purpose of this paper is to prove that token-based inpainting is promising for visual reconstruction.

3.3. Hierarchical Reconstruction Architecture

We find that, for fMRI decoding, many instances almost invariably recover outline information first after several epochs, and overfitting phenomenon occurs when decoding intricate color textures, as shown in Fig. 3.

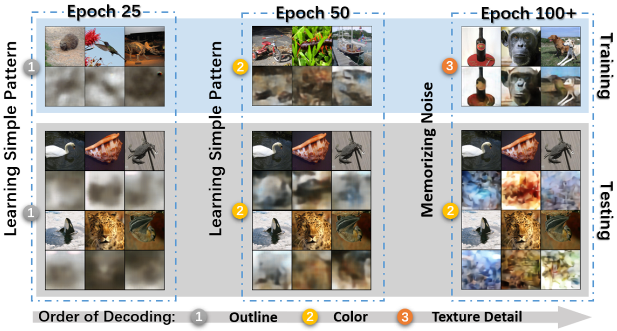

Figure 3. Some phenomena that occur during visual decoding.

Given the above observation and the analysis in (Arpit et al., 2017), there was a logical explanation. Because neural networks tend to preferentially learning simple patterns, and low-frequency information is relatively robust to noise in the process of visual encoding. Thus, its patterns are captured early in the training process. In turn, the reconstruction network may be brute-force memorizing noise, which leads to overfitting, when trying to decode high-frequency information, and this is even more severe for limited training data. Early stopping strategy or regularization can alleviate this phenomenon, but high-quality images require more prototype vectors to reconstruct, which also increases the difficulty of establishing visual cues. Thus, it is necessary to avoid using high-frequency signals to supervise fMRI representation learning. Toward this end, we provide a hierarchical model (Fig. 4), i.e., compression followed by a super-resolution strategy, to alleviate the need to directly predict image texture details from brain activity signals.

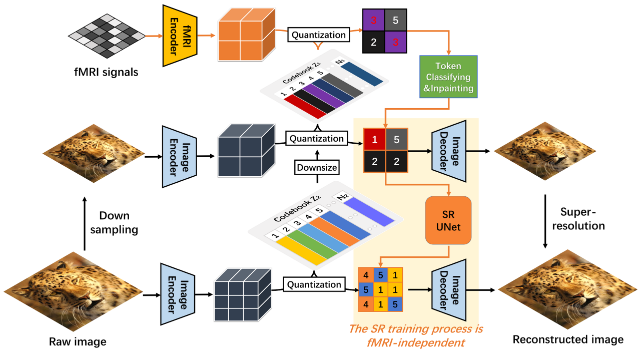

Figure 4. The proposed super-resolution (SR) architecture. The training phase is divided into two steps: learning a set of multiscale codebooks, and then searching the mapping between the image tokens composed of different codebooks.

In practice, the size of the codebook and the downsampling factor of the encoder in VQ-VAE determine the image reconstruction capability, which enables us to remove the high-frequency signature by reducing image resolution and codebook size. Hence, the final architecture, customized as a hierarchical structure, for learning two image codebooks at different scales. First, the utilization of a small codebook serves as a guide for building visual cues, effectively reducing the complexity of image features and spatial structures, which is equivalent to weak the difficulty of fMRI embedding and token-to-token repair learning. We then leverage UNets (Ronneberger et al., 2015) to learn the mapping relation between multi-scale image tokens, and the image super-resolution is fulfilled by the collaborative use of decoder of VQ-VAE, which can be written as:

$$\mathcal{Z}^y_{sr} = \mathbf{VQ}\Big(\mathbf{F}_{sr}\Big(\mathbf{VQ}\Big(\mathbf{E}_L(y\downarrow); Z_L\Big)\Big)\Big), \tag{7}$$

$$s.t.\ \mathbf{VQ}\Big(\mathbf{E}(y); Z\Big) = \mathcal{Z}^y_{sr}, \tag{8}$$

$$y^* = \mathbf{D}(\mathcal{Z}^y_{sr}). \tag{9}$$

where ↓ denotes the downsampling operation, Z_L is the pretrained small scale codebook, and $\mathbf{F}_{sr}$, modeled by UNet, would be viewed as a transfer function of codebook entries.

The target loss function $\mathcal{L}_{sr}$ is

$$\mathcal{L}_{sr} = \Big|\Big|y^* - \mathbf{D}(\mathcal{Z}^y)\Big|\Big|_2^2 + \mathcal{L}_{vm}\Big(\mathcal{Z}^y_{sr}, \mathbf{VQ}\Big(\mathbf{E}(y); Z\Big)\Big). \tag{10}$$

Note that the training step of super-resolution does not deal with any fMRI data. Finally, given a fMRI example x, its visual decoding process can be formulated as:

$$\mathcal{Z}^x_L = \mathbf{VQ}\Big(\mathbf{UNet}\Big(\mathrm{MLP}(x)\Big); Z_L\Big), \tag{11}$$

$$\mathcal{Z}^x_{sr} = \mathbf{F}_{sr}\Big(\mathbf{F}_{token}(\mathcal{Z}^x_L)\Big), \tag{12}$$

$$y^* = \mathbf{D}(\mathcal{Z}^x_{sr}), \tag{13}$$

where $\mathbf{F}_{token}(\cdot)$ is token-to-token inpainting function, described in Sec. 3, and $\mathbf{D}(\cdot)$ indicates the decoder of VQ-VAE with codebook Z.

4. Experimental Results

Experiments were carried out with a benchmark dataset in comparison with existing approaches.

4.1. Dataset and Settings

Benchmark Dataset. We experimented with a popular publicly available fMRI dataset, which is called Generic Object Decoding (GOD) dataset (Horikawa & Kamitani, 2017). The dataset provides stimulus images and the evoked fMRI recordings, where visual images are selected from ImageNet, and presented with fixation in a 3T scanner (TR, 3s; voxel size, 3×3×3 mm). Specifically, five subjects were presented with 500×500 color images from 150 categories and taken the related visual regions of interest in brain (including V1-V4, LOC, FFA and PPA). During the training phase, we follow the original training/test set split. For each subject, training set consists 1200 fMRI-image pairs, and the testing made up of 50 fMRI recordings with corresponding images.

Details of Implementation. The parameter setting of VQ-fMRI for all experiments is summarized as follows. Enocders of VQ-VAE: 2 convolutional layers (stride 2, kernel 4×4, and padding 1), followed by two residual blocks; Deocders of VQ-VAE: two residual blocks, followed by 3 transposed convolutions (stride 2, kernel 4×4, and padding 1); Codebooks: $Z_L \in \mathbb{R}^{8\times32}$ (image $y \in \mathbb{R}^{64\times64\times3}$), and $Z \in \mathbb{R}^{8\times128}$ (image $y \in \mathbb{R}^{128\times128\times3}$). We implemented the image classifier, inpainting, and SR modules using the UNet with 2 downsampling and 2 upsampling layers (stride 2, kernel 4×4, and padding 1). Adam solver (Kingma & Ba, 2014) is employed to optimize the parameters with a learning rate of 2e-4. We pre-train the VQ-VAEs, token inpainting, and SR modules on the ImageNet dataset (Deng et al., 2009). All competitors are implemented based on official codes with the optimal parameter settings.

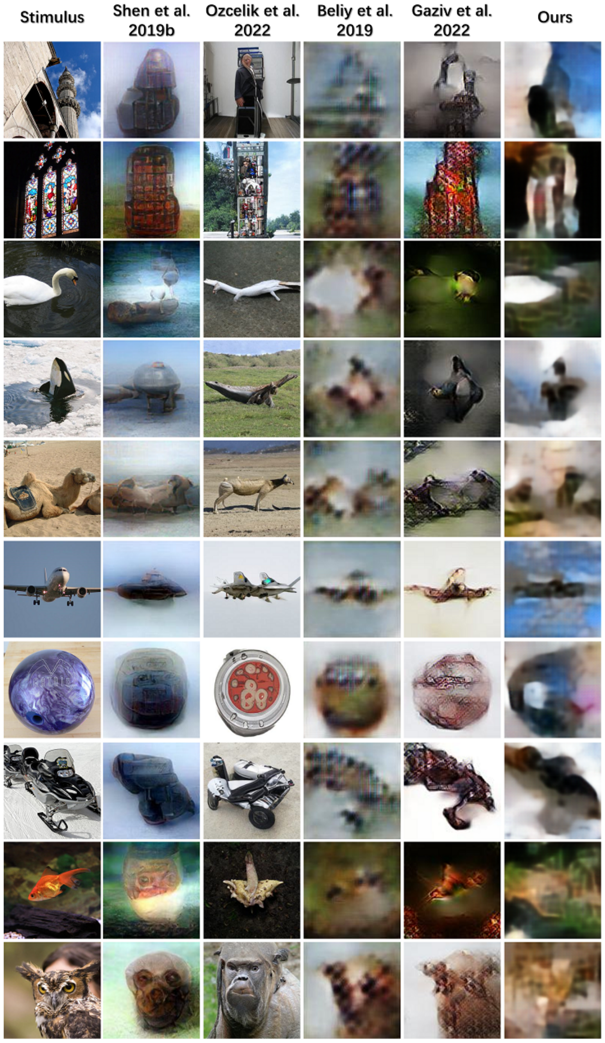

Figure 5. Comparison of reconstruction results for our VQ-fMRI with four state-of-the-art fMRI decoding methods. The first column is the real stimulus images.

4.2. Evaluation Metric

Following (Rakhimberdina et al., 2021; Shen et al., 2019b), we leverage two image evaluation settings for quantitative comparison: 1) one-to-one, and 2) pairwise comparison.

One-to-one Evaluation. It evaluates the similarity score between the reconstruction and the ground truth via a specific metric. In our experiments, structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and pixel-wise Pearson correlation coefficient (PCC) are used.

Pairwise Evaluation. It is performed by comparing a reconstructed visual image with two candidate images (including ground truth and a randomly selected image). If the metric score involving the real image is better than that with the non-relevant image, we consider that the trial is correct. PCC is a popular metric in pairwise comparison (Rakhimberdina et al., 2021), and we also follow such practice.

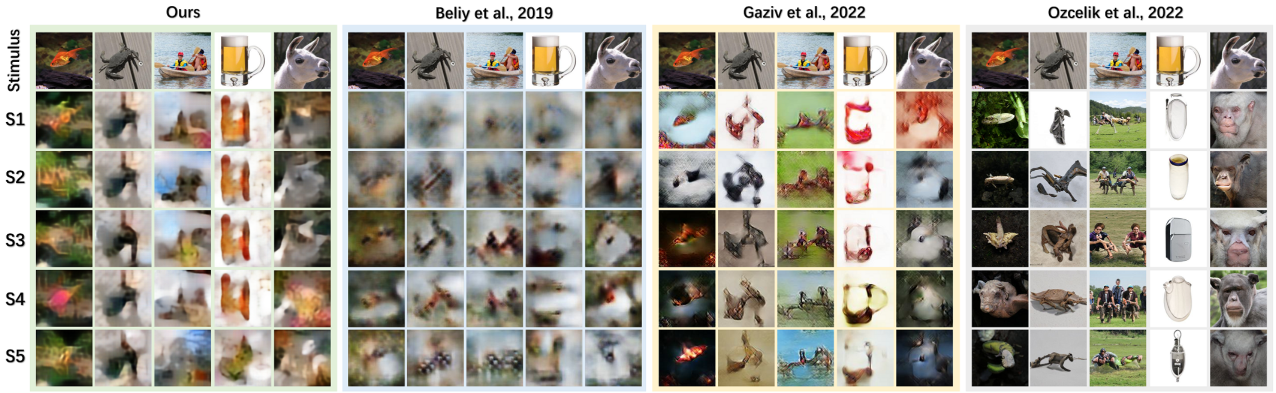

Figure 6. Reconstruction results of the four methods on all subjects (S1-S5). For each group, the first row provides ground truth images.

4.3. Comparison with State-of-the-Art

Here we evaluate the reconstructed images in comparison with existing approaches. The competitors are four representative approaches of (Shen et al., 2019b), (Beliy et al., 2019), (Gaziv et al., 2022) and (Ozcelik et al., 2022), including two encoder-decoder-based approaches ((Beliy et al., 2019) and (Gaziv et al., 2022)) and two GAN-based approaches ((Shen et al., 2019b) and (Ozcelik et al., 2022)), representing the state-of-the-art.

Table 1. Quantitative comparison of five methods (↑ indicates the higher the better). **Bold** represents the optimal indicator value.

Method	SSIM ↑	PSNR ↑	PCC ↑
(Shen et al., 2019b)	0.413±0.154	10.7±1.97	0.482±0.176
(Ozcelik et al., 2022)	0.385±0.163	10.0±2.42	0.241±0.131
(Beliy et al., 2019)	0.432±0.162	12.2±2.39	0.429±0.149
(Gaziv et al., 2022)	0.372±0.155	10.3±2.91	0.424±0.164
Ours	**0.492±0.125**	**13.4±1.76**	**0.551±0.122**

Firstly, we present the reconstruction images recovered by VQ-fMRI in comparison with existing approaches. In Fig. 5, the first column is the original stimulus image, and the reconstruction results are illustrated in the rest columns. From the intuitive visual results, we can perceive that VQ-fMRI can successfully reconstruct shapes, color details, and global layouts. Compared with encoder-decoder approaches of (Beliy et al., 2019) and (Gaziv et al., 2022), VQ-fMRI demonstrates a higher capacity to recover the color information consistent with the stimulus image in most cases, for instance, the blue sky in the 1^{st} and 6^{th} samples. GAN-based approaches ((Shen et al., 2019b) and (Ozcelik et al., 2022)) obtain images with higher quality, but the reconstructions are somehow deviant from the real visual stimulus. In contrast, the reconstructions with VQ-fMRI exhibit a more consistent layout and content of the images.

Then we quantitatively compare the visual reconstruction performance with one-to-one settings, using the SSIM, P-SNR, and PCC criteria, and the results are shown in Tab. 1. Overall, VQ-fMRI outperforms the competitors with all three criteria. Specifically, with the SSIM, which reflects the similarity of local spatial pixels, VQ-fMRI obtains a high value of 0.49, which is 18% to 32% higher than the competitors. With the PCC, which computes the linear relationship between two image variables, our method achieves a value of 0.55, which is 14% to 128% higher than the competitors. It is notable that, although GAN-based methods such as (Ozcelik et al., 2022) generate pleasant natural appearances, the reconstructions are usually deviant from the stimulus, such that obtain lower SSIM and PCCs. These two criteria indicate that the images reconstructed with VQ-fMRI faithfully reveal the stimulus images. VQ-fMRI also reaches a high PSNR (dB as unit), which is 9% to 33% higher than other competitors, indicating the reconstructed images of VQ-fMRI can better preserve the raw visual structures.

4.4. Comparison with Different Subjects

Comparing across different subjects demonstrates the robustness of an approach (Rakhimberdina et al., 2021). Therefore, to provide a comprehensive quantitative evaluation of our VQ-fMRI in dealing with different subjects, we conduct experiments on five subjects of the GOD dataset, and compare with three competitors of (Beliy et al., 2019), (Gaziv et al., 2022) and (Ozcelik et al., 2022).

We compare the reconstructed images across different subjects in Fig. 6. The first row represents the original stimulus image, and the rest rows are the reconstruction using fMRI from different subjects. We can see there existed individual-wise bad cases, such as the "cup" for S5 (Fig. 6, row 6 col 4). As it is only a bad case with the S5, we think it was due to low quality in fMRI signals or the subject was not fully focused when viewing the image. On the whole, however, our approach obtains superior performance with consistent

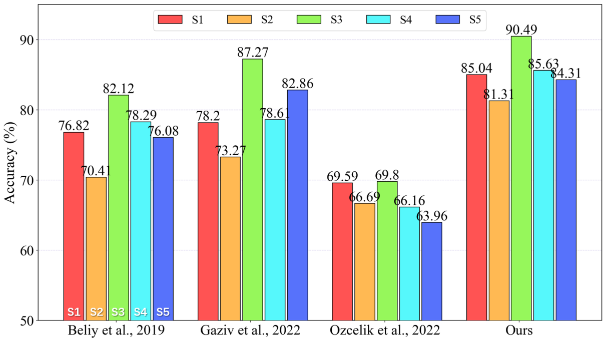

Figure 7. Pairwise PCC evaluation across all subjects.

reconstructions with different subjects. With the constraint of discrete cues conditions, the reconstructed images across different subjects share similar structures with the raw visual stimulus, which benefits from the inpainting model that is regulated by the distribution of natural image tokens. These intuitionistic observations are supported by the quantitative comparison in Tab. 2 and Fig. 7.

Compared with the competitors, since (Ozcelik et al., 2022) aimed at matching higher-level semantics, it does not perform favorably on low-level image measures. On the other hand, although both (Beliy et al., 2019) and (Gaziv et al., 2022) are self-supervised methods, concatenating encoder and decoder back-to-back, which require separate training on unlabeled fMRI and image data, makes it more susceptible to catastrophic forgetting problems. Conversely, bridging multi-modal data with discrete codebooks can effectively prevent that. Results show that our VQ-fMRI achieves a lead of 1.45% to 8.04% on the pairwise PCC.

4.5. Ablation Study

Here, we first study the impact of VQ-VAE architecture settings (including codebook and token sequence) on decoding performance (see Tab. 3), and then evaluate the effectiveness of the core components of the proposed VQ-fMRI framework: the inpainting process, and the super-resolution process (presented in Tab. 4 and Fig. 8). Finally, we also provide further evaluation into the robustness of the presented model via different folds of training/test data (summarized in Tab. 5). In this experiments, we repeat the quantitative comparison with subject 3.

Codebook and Token Sequence. The ablation results (see Tab. 3) show that the codebook size and the token sequence length had a direct effect on the quality of the recovered images. We observe that even if we use a small codebook size (i.e., K = 8, and d = 32), the reconstructions look only slightly blurrier than the originals, and the overall visual structure of images can be preserved. On the other hand, with a high compression ratio (i.e., h = 8, and w = 8), the quality of the reconstructed images can be improved. We also notice that a too-small codebook size and feature map could lead to performance degradation.

Table 2. One-to-one evaluation across all subjects for four methods. The optimal indicator values are presented in **bold**.

Sub	Method	SSIM ↑	PSNR ↑	PCC ↑
S1	(Beliy et al., 2019)	0.383±0.122	11.0±2.27	0.330±0.194
	(Gaziv et al., 2022)	0.324±0.144	8.94±2.79	0.302±0.184
	(Ozcelik et al., 2022)	0.338±0.151	9.27±**2.23**	0.181±**0.145**
	Ours	**0.414±0.122**	**12.8**±2.37	**0.372**±0.179
S2	(Beliy et al., 2019)	0.366±**0.113**	11.1±**2.08**	0.303±0.200
	(Gaziv et al., 2022)	0.309±0.137	9.42±2.54	0.297±0.202
	(Ozcelik et al., 2022)	0.345±0.148	9.85±2.30	0.220±**0.171**
	Ours	**0.407**±0.133	**12.7**±2.66	**0.364**±0.253
S3	(Beliy et al., 2019)	0.382±0.123	11.7±1.96	0.356±0.191
	(Gaziv et al., 2022)	0.347±0.138	10.7±2.47	0.380±0.186
	(Ozcelik et al., 2022)	0.352±0.160	10.3±**1.92**	0.225±**0.153**
	Ours	**0.423±0.114**	**13.2**±1.92	**0.419**±0.193
S4	(Beliy et al., 2019)	0.357±**0.112**	10.9±2.07	0.314±0.200
	(Gaziv et al., 2022)	0.346±0.126	11.0±2.21	0.315±0.192
	(Ozcelik et al., 2022)	0.340±0.151	9.78±2.36	0.232±**0.187**
	Ours	**0.378**±0.114	**13.2±2.04**	**0.352**±0.195
S5	(Beliy et al., 2019)	0.353±**0.121**	10.6±2.24	0.289±0.209
	(Gaziv et al., 2022)	0.342±0.139	10.6±2.48	0.314±0.206
	(Ozcelik et al., 2022)	0.351±0.156	9.90±**2.03**	0.207±**0.156**
	Ours	**0.380**±0.126	**12.8**±2.18	**0.348**±0.205

Table 3. Influence of the parameters of VQ-fMRI. We report the pairwise comparison accuracy of SSIM, PSNR, and PCC on GOD.

Codebook	Token Sequence	SSIM ↑	PSNR ↑	PCC ↑
K = 4, d = 16	h = 4, w = 4	68.01%	73.95%	87.46%
	h = 8, w = 8	69.11%	78.29%	88.49%
	h = 16, w = 16	65.36%	70.97%	84.34%
K = 8, d = 32	h = 4, w = 4	70.54%	79.35%	90.01%
	h = 8, w = 8	71.35%	81.73%	90.49%
	h = 16, w = 16	66.23%	72.36%	83.36%
K = 16, d = 64	h = 4, w = 4	69.32%	77.71%	89.60%
	h = 8, w = 8	70.62%	81.99%	88.63%
	h = 16, w = 16	65.24%	71.35%	82.64%

Inpainting. The token-based inpainting process, which completes the images with experience-based content completion, brings a substantial performance boost to the reconstruction. Specifically, by using the inpainting step, the pairwise SSIM, PSNR, and PCC increase by 8.8%, 5.8%, and 15.6% respectively, as shown in Tab. 4. Comparing the reconstructed images with and without inpainting (the 3^{rd} and 4^{th} rows in Fig. 8), we find that, the inpainting process help enrich the color and details of the images effectively.

Especially, the border between the foreground and background is more clear, and the layout and colors are more accurate. While there are also bad cases such as in the last sample in Fig. 8, which may be caused by the ambiguity of the discrete cues.

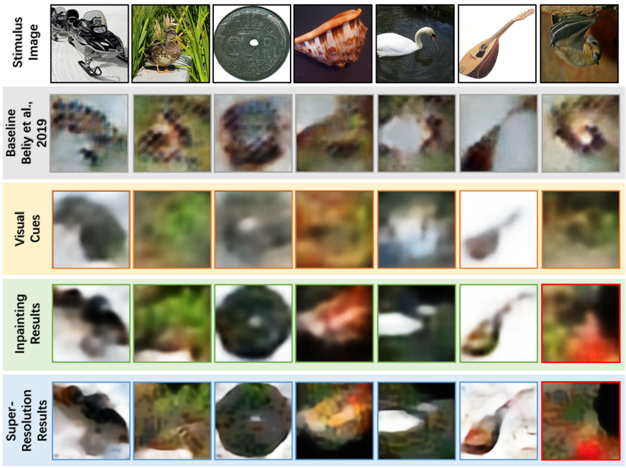

Figure 8. Several intuitive results of ablation study. The first and second rows show the ground truth images, the baselines, and the reconstructed results from the fMRI visual cues, respectively.

Super-Resolution. The super-resolution process helps further improve the quality of the image. Overall, the super-resolution further improves the performance of the pairwise SSIM, PSNR, and PCC by 0.27%, 1.55%, and 0.84% respectively. While the performance gain brought by SR is marginal compared with the inpainting process. It could be because the SR process can also cause biases, especially for some image-wise bad cases such as the "bat" (Fig. 8, col 7). Note that image-wise bad case means that the reconstructed images for all subjects has obvious decoding deviations. We think this type of bad case might be due to the prior knowledge in the inpainting model did not well cover such images, which may be a limitation of this work. Nevertheless, most of the SR results demonstrate high-quality reconstructions, which may provide an interesting topic in the interdisciplinary study of brain signal decoding and CV.

Table 4. Reconstruction performance, evaluated by three pairwise similarity metrics, on different architectures of our VQ-fMRI.

Inpaint	SR	**SSIM ↑**	**PSNR ↑**	**PCC ↑**
×	×	62.24%	74.33%	74.00%
✓	×	71.08%	80.18%	89.65%
✓	✓	71.35%	81.73%	90.49%
baseline (Beliy et al., 2019)		67.84%	73.27%	82.12%

Evaluation of the Robustness. Finally, we test the robustness of our VQ-fMRI with different folds of training/test data. Specifically, instead of using a fixed test set, we randomly sampled 50 fMRI-image pairs as the test set, and the results are averaged with five independent runs, as reported in Tab. 5. The performance of original training/test split is used as a baseline. We see that pairwise SSIM metric shows the largest decline compared to the baseline, but the influence of changing training/test data on reconstruction performance is relatively limited (smaller than 2.26%).

Table 5. Robustness test of VQ-fMRI on different training/test splits. Note that the parenthetical value denotes std.

Run No.	**SSIM ↑**	**PSNR ↑**	**PCC ↑**
1	69.12% (0.34)	80.92% (0.26)	88.85% (0.23)
2	68.85% (0.36)	80.70% (0.23)	86.74% (0.25)
3	67.77% (0.35)	79.02% (0.22)	87.36% (0.23)
4	69.08% (0.32)	80.47% (0.25)	89.05% (0.22)
5	68.36% (0.33)	78.79% (0.20)	86.98% (0.24)
baseline	71.35% (0.32)	81.73% (0.20)	90.49% (0.22)
mean (std)	69.09% (0.35)	80.27% (0.24)	88.25% (0.24)

5. Conclusion

In this paper, we present a novel fMRI-to-image transform architecture, named VQ-fMRI, to revisit visual neural decoding. Unlike existing GAN-based and diffusion-based models that focus on recovering high-quality semantically correct images, this work makes efforts to reconstruct overall visual organization of seen images at the pixel level. For this purpose, the core idea is to imitate the way of looking at the world in our brain, rather than following existing popular paradigm (i.e., seeking the strong equivalence of neural representations). The proposed token-to-token inpainting and super-resolution strategy avoid to decode imperceptible feature details from fMRI data, thus effectively reducing the reconstruction errors, and guaranteeing that generated images are semantically meaningful. In general, our model has the capacity to generate images that are more in line with the actual visual stimuli, and surpasses leading alternatives. The principle of our method is general, which is expected to be popularized to other neural decoding fields (e.g., audio decoding).

Acknowledgements

This work was supported in part by the Science and Technology Innovation 2030 Major Projects (2021ZD0200400), the Key Research and Development Program of Zhejiang Province in China (2020C03004), the Natural Science Foundation of China (NSFC) (U1909202, 61925603, 62276228), and the Lingang Laboratory (LG-QS-202202-04).

References

Arpit, D., Jastrzundefinedbski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pp. 233–242. PMLR, 2017.

Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., and Irani, M. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Casanova, A., Careil, M., Verbeek, J., Drozdzal, M., and Romero Soriano, A. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

Chen, Z., Qing, J., Xiang, T., Yue, W. L., and Zhou, J. H. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22710–22720, June 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Desimone, R., Duncan, J., et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

Du, C., Du, C., Huang, L., and He, H. Reconstructing perceived images from human brain activities with bayesian deep multiview learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8):2310–2323, 2018.

Du, C., Du, C., Huang, L., Wang, H., and He, H. Structured neural decoding with multitask transfer learning of deep neural network representations. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):600–614, 2022.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Fang, T., Qi, Y., and Pan, G. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33: 13038–13048, 2020.

Gaziv, G., Beliy, R., Granot, N., Hoogi, A., Strappini, F., Golan, T., and Irani, M. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022.

Gazzaniga, M. S. *The cognitive neurosciences*. MIT press, 2009.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 26722680. MIT Press, 2014.

Hinton, G. E. and Lang, K. J. Shape recognition and illusory conjunctions. In *International Joint Conference on Artificial Intelligence*, pp. 252–259, 1985.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Horikawa, T. and Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):1–15, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., Mack, S., et al. *Principles of Neural Science*, volume 4. McGraw-hill New York, 2000.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Liu, H., Jiang, B., Song, Y., Huang, W., and Yang, C. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, pp. 725–741. Springer, 2020.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

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

Mozafari, M., Reddy, L., and VanRullen, R. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L., and VanRullen, R. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022.

Qian, C., Sun, X., Wang, Y., Zheng, X., Wang, Y., and Pan, G. Binless kernel machine: Modeling spike train transformation for cognitive neural-prostheses. *Neural Computation*, 32(10):1863–1900, October 2020.

Qiao, K., Chen, J., Wang, L., Zhang, C., Tong, L., and Yan, B. Biggan-based bayesian reconstruction of natural images from human brain activity. *Neuroscience*, 444: 92–105, 2020.

Rakhimberdina, Z., Jodelet, Q., Liu, X., and Murata, T. Natural image reconstruction from fmri using deep learning: A survey. *Frontiers in Neuroscience*, 15:795488, 2021.

Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., and Gao, X. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, June 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

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

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., and Kamitani, Y. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, pp. 21, 2019a.

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):1–23, 01 2019b.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

St-Yves, G. and Naselaris, T. Generative adversarial networks conditioned on brain activity reconstruct seen images. In *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1054–1061. IEEE, 2018.

Takagi, Y. and Nishimoto, S. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, June 2023.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Wang, Y., Lu, M., Wu, Z., Tian, L., Xu, K., Zheng, X., and Pan, G. Visual cue-guided rat cyborg for automatic navigation. *IEEE Computational Intelligence Magazine*, 10(2):42–52, May 2015.

Wu, Z., Pan, G., and Zheng, N. Cyborg intelligence. *IEEE Intelligent Systems*, 28(5):31–33, 2013.

Yu, Y., Pan, G., Gong, Y., Xu, K., Zheng, N., Hua, W., Zheng, X., and Wu, Z. Intelligence-augmented rat cyborgs in maze solving. *PloS One*, 11(2):e0147754, 2016.