A Review on Reconstruction of Images from functional MRI

Muhammad Abdul Rafey Farooqi, Hashir M. Kiani

December 6, 2025

Abstract

The reconstruction of visual stimuli from human brain activity represents a fundamental challenge in computational neuroscience, offering a window into the internal representations of the mind. This review provides a comprehensive analysis of the field, beginning with the biological underpinnings of the visual cortex and the signal characteristics of functional Magnetic Resonance Imaging (fMRI). We mathematically formulate the reconstruction task as an ill-posed inverse problem, where the objective is to recover high-dimensional visual information from sparse, noisy hemodynamic signals. The review traces the methodological evolution of the field, from early linear decoders and hand-crafted feature bases to the current paradigm shift driven by Deep Neural Networks (DNNs) and generative artificial intelligence. Special attention is given to state-of-the-art approaches utilizing Denoising Diffusion Probabilistic Models (DDPMs) and Vision-Language Models (VLMs), which have achieved unprecedented photorealism by effectively leveraging strong semantic priors. Finally, we critically examine the limitations of current techniques – including cross-subject generalization, temporal resolution, and generative hallucinations – and discuss the benchmark datasets that are shaping the future of neural decoding.

Introduction

This project focuses on visual reconstruction: the task of decoding and regenerating the visual stimuli viewed by a subject solely from their functional Magnetic Resonance Imaging (fMRI) recordings.

This area of research is of critical importance for the development of non-invasive Brain-Computer Interfaces (BCIs). Successful reconstruction technology could eventually enable direct communication for individuals with speech or motor impairments (such as those with Locked-in Syndrome) by translating mental imagery directly into digital outputs. However, this task presents a significant challenge: fMRI signals are noisy, indirect measures of neural activity, making the mapping from brain signals back to images a highly ill-posed inverse problem.

The primary goal of this project is to determine how noisy, low-dimensional fMRI signals can be effectively mapped to the latent space of a generative diffusion model to reconstruct high-fidelity images that are both visually accurate and semantically consistent with the original stimulus.

Related Work

The human visual system is organized hierarchically. Visual information enters the retina and is transmitted via the lateral geniculate nucleus (LGN) to the primary visual cortex (V1) located in the occipital lobe. Early visual cortex areas (V1–V3) are retinotopically organized, meaning adjacent neurons process adjacent regions of the visual field. V1 neurons act as local edge detectors, sensitive to orientation and spatial frequency, functionally resembling Gabor filters.

As information progresses along the ventral stream ("what" pathway; see Figure 1), representations become increasingly abstract and invariant to position, encoding complex shapes, objects, and faces rather than simple pixel-level details. This biological hierarchy ($V1 \rightarrow$ Object) has inspired the use of Deep Neural Networks (DNNs) in reconstruction, as CNNs exhibit a similar feature hierarchy.

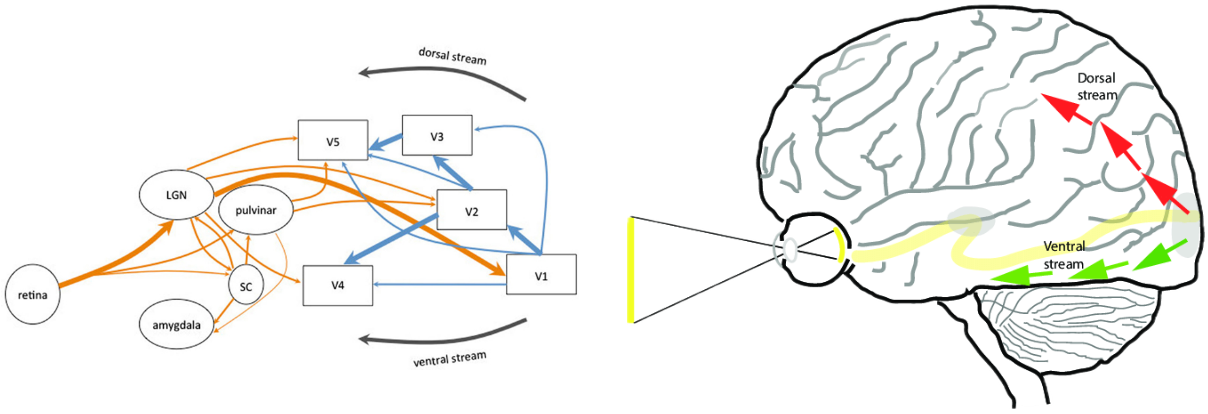

Figure 1: Schematic overview of the early and higher visual pathways, illustrating the flow of information from the retina through LGN to cortical areas involved in feature extraction and object processing. Referenced from [15] and [16]

Functional Magnetic Resonance Imaging (fMRI) is the primary non-invasive tool for measuring brain activity. However, it does not measure neuronal firing directly; instead, it measures the Blood Oxygen Level Dependent (BOLD) signal. When neurons are active, they consume oxygen, and the body overcompensates by increasing blood flow to that region, changing the ratio of oxygenated (diamagnetic) to deoxygenated (paramagnetic) hemoglobin. This magnetic change is detected by the MRI scanner. Crucially, this process is modeled by the Hemodynamic Response Function (HRF), a temporal filter that introduces a delay (peaking ≈ 6 seconds after stimulus) and acts as a low-pass filter on the neural signal.

Mathematically, we frame visual reconstruction as an inverse problem. Let $\mathbf{x} \in \mathbb{R}^N$ represent the visual stimulus (an image) and $\mathbf{y} \in \mathbb{R}^V$ represent the recorded BOLD signal across V voxels. The forward process (encoding) is modeled as

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \epsilon, \tag{1}$$

where $\mathcal{F}$ is a composite operator representing the visual system's transformation of light into neural spikes, convolved with the HRF, and ϵ represents noise inherent in fMRI acquisition. The goal is to find the inverse mapping $\hat{\mathbf{x}} = \mathcal{F}^{-1}(\mathbf{y})$, which is ill-posed because the mapping is many-to-one and $N \gg V$.

To solve this, we employ a probabilistic framework:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} P(\mathbf{y} \mid \mathbf{x}) P(\mathbf{x}), \tag{2}$$

where $P(\mathbf{x})$ is a prior constraining the solution to resemble a natural image (Figure 2).

Initial approaches assumed a linear relationship between pixel contrast and voxel activity. Kay et al. (2008) [1] utilized Gabor wavelet pyramids to model the receptive fields of V1 voxels, allowing them to identify which natural image a subject was viewing from a closed set. Miyawaki et al. (2008) [2] achieved true reconstruction by treating the image as a 10×10 mosaic. They trained a "multi-scale local image decoder" – essentially a linear regression model – to predict the contrast of each patch. While effective for simple shapes (squares, letters), this method failed to capture the complexity of natural scenes.

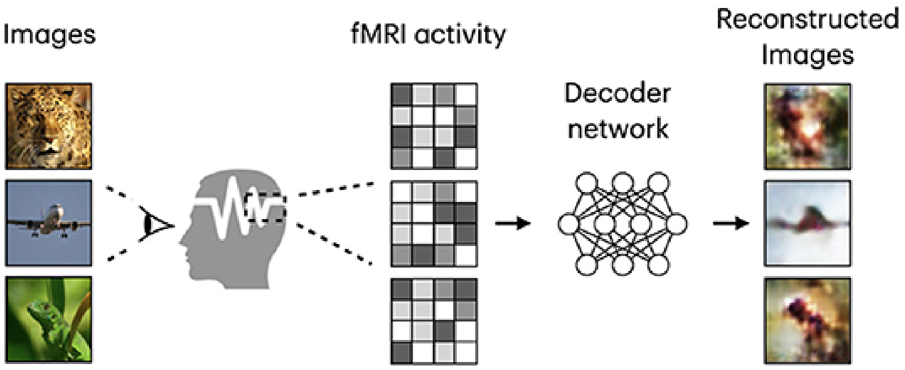

Figure 2: Illustration of the visual reconstruction pipeline. Left: the forward (encoding) process maps a visual stimulus x through sensory and physiological transformations into a BOLD fMRI response $y = F(x) + \varepsilon$. Right: the inverse (decoding) process seeks a reconstruction $\hat{x} = \arg\max_x P(y \mid x)\, P(x)$, where $P(x)$ is a prior enforcing natural-image plausibility. Referenced from [17]

The field shifted significantly with the introduction of Deep Neural Networks (DNNs). Horikawa and Kamitani (2017) [3] demonstrated that fMRI activity could be translated into the feature space of a Convolutional Neural Network (e.g., AlexNet). They showed that lower visual areas (V1–V3) predicted early CNN layers, while higher areas predicted deep layers. Shen et al. (2019) [4] utilized this for reconstruction, freezing a pre-trained CNN and optimizing an input image $\mathbf{x}$ such that its CNN features matched those predicted from the brain activity:

$$\min_{\mathbf{x}} \|\phi_{\text{CNN}}(\mathbf{x}) - \hat{\phi}_{\text{brain}}(\mathbf{y})\|^2. \tag{3}$$

This produced recognizable objects, but images were often blurry or "ghostly." To improve realism, researchers introduced Generative Adversarial Networks (GANs) as a learnable prior [5], which forced the output to look realistic, sometimes at the cost of semantic accuracy.

The introduction of Denoising Diffusion Probabilistic Models (DDPMs) and Vision-Language Models (VLMs) like CLIP revolutionized the field. Takagi and Nishimoto (2023) [6] and Chen et al. (2023) (MinD-Vis) proposed "Latent Diffusion Decoding." Instead of predicting pixels, they trained linear models to map fMRI activity to the latent space of Stable Diffusion (variational autoencoder latent z and text condition c). Injecting these predicted latents into the denoising process generated high-resolution, photorealistic images, with the brain signal effectively "guiding" the diffusion model to match neural data.

The current benchmarks in neural reconstruction are held by MindEye [8] and MindEye2 [9] as shown in Figure 3. These studies highlight that accurate reconstruction requires capturing both high-level semantic information and low-level perceptual structure from fMRI signals. The semantic component maps fMRI responses to the CLIP text embedding space, encoding conceptual information (e.g., "a red car"), while the perceptual component maps fMRI to a low-level visual embedding that preserves spatial structure, shape, and color. By aligning voxels to these dual embedding spaces using a large Transformer-based backbone (MLP-Mixer) and conditioning a diffusion model on both, these methods produce reconstructions that are simultaneously semantically meaningful and structurally faithful, achieving state-of-the-art performance with as little as one hour of training data per subject.

Datasets and Benchmarking

The development of robust reconstruction models has been inextricably linked to the availability of large-scale, high-quality public datasets. Unlike computer vision, where models are trained

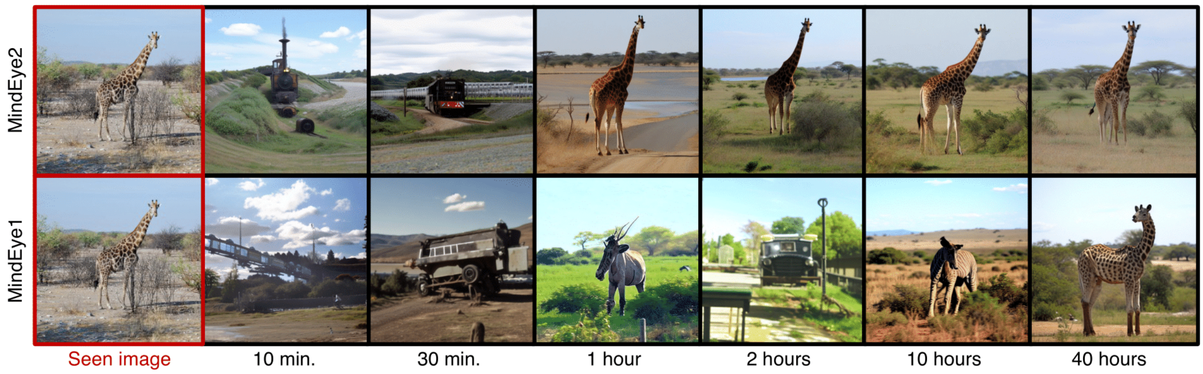

Figure 3: MindEye2 vs. MindEye1 reconstructions from fMRI brain activity using varying amounts of training data. Referenced from [9]

on millions of images, fMRI research is constrained by the physiological limits of human subjects and the immense cost of scanning. Consequently, the field relies on a few landmark datasets that serve as common benchmarks for comparing model performance.

For nearly a decade, the primary benchmark was the Generic Object Decoding dataset introduced by Horikawa and Kamitani. This dataset consists of fMRI responses from five subjects viewing 1,200 training images and 50 test images drawn from ImageNet. While foundational, its relatively small size limited the complexity of the models that could be trained without overfitting. To address the need for greater scale and diversity, Chang et al. (2019) [13] introduced BOLD5000, a dataset comprising roughly 5,000 images from standard computer vision datasets like COCO and ImageNet. This initiative explicitly aimed to bridge the gap between human vision and computer vision, allowing researchers to test whether models trained on standard vision tasks could generalize to brain data.

However, the current era of generative reconstruction was largely catalyzed by the release of the Natural Scenes Dataset (NSD) by Allen et al. (2022) [14]. The NSD is currently the largest, high-resolution fMRI dataset available, containing responses from eight subjects who each viewed up to 73,000 distinct natural scenes from the COCO dataset over the course of a year. The sheer volume of this data – roughly an order of magnitude larger than previous collections – enabled the training of data-hungry architectures like Transformers and diffusion models (e.g., MindEye), which require massive amounts of training examples to learn the subtle mapping between voxel patterns and visual semantics.

Evaluating these models requires a diverse set of metrics, as no single score captures the multifaceted nature of visual perception. Early studies relied heavily on pixel-wise metrics such as Mean Squared Error (MSE) and Pearson correlation. However, these metrics often penalize realistic reconstructions that are slightly misaligned spatially, while favoring blurry, averaged images that minimize variance. To capture structural fidelity, the field adopted the Structural Similarity Index (SSIM), which assesses the preservation of luminance, contrast, and structure.

In the current generative era, benchmarking has shifted towards semantic evaluation. Researchers now routinely employ N-way identification tasks, where a pre-trained classifier (such as a ResNet or CLIP model) determines whether the reconstructed image is closer to the ground truth than to a set of distractors. Furthermore, metrics like CLIP-similarity are used to quantify the semantic distance between the text description of the original image and the reconstruction, ensuring that the model captures the conceptual content (e.g., “a dog on a beach”) even if the exact pixel arrangement differs. This combination of low-level structural metrics and high-level semantic metrics provides a holistic view of model performance.

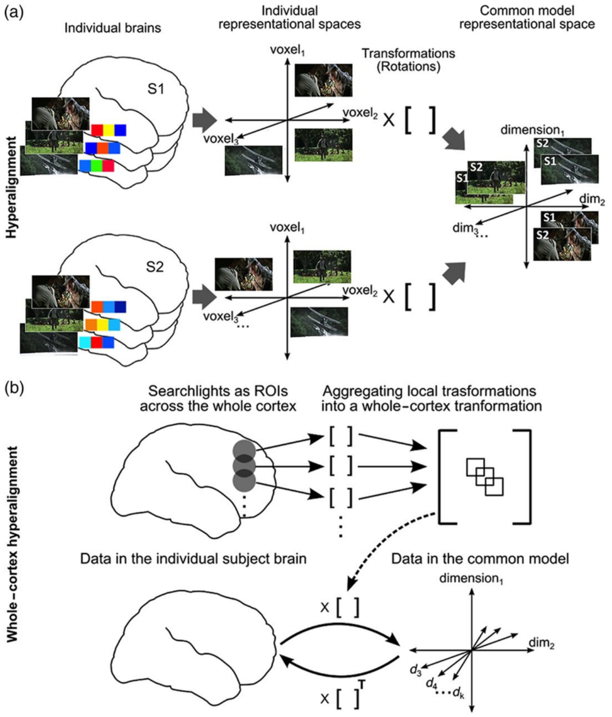
(a)
Individual brains
Individual representational spaces
Common model representational space
Transformations (Rotations)
Hyperalignment
S1
S2
voxel₁
voxel₂ x [ ]
voxel₃
dimension₁
dim₂
dim₃...
(b)
Searchlights as ROIs across the whole cortex
Aggregating local trasformations into a whole-cortex tranformation
Whole-cortex hyperalignment
Data in the individual subject brain
Data in the common model
dimension₁
x [ ]
dim₂
d₃ d₄ ... dₖ
x [ ]ᵀ

Figure 4: Schematic of whole-cortex searchlight hyperalignment (common model of representational spaces). Each individual subject's voxel space is transformed via an orthogonal rotation into a shared high-dimensional representational space. This enables alignment across subjects despite anatomical and functional variability, facilitating universal decoding of fMRI signals. Referenced from [18]

Gaps and Future Opportunities

Despite the rapid progress driven by generative AI, several critical challenges remain before fMRI reconstruction can be deployed in practical or clinical settings. A primary obstacle is Cross-Subject Generalization. Current state-of-the-art models are almost exclusively subject-specific; a model trained on one participant typically fails when tested on another due to the high variability in functional brain organization and cortical anatomy. Training a high-performance decoder currently requires collecting massive datasets – often 15 to 40 hours of scan time from a single individual – which is prohibitively expensive for general populations. A promising avenue to address this is the development of "Universal Decoders" via functional alignment techniques. Haxby et al. (2011) [10] pioneered "Hyperalignment," a method to rotate the high-dimensional voxel spaces of different subjects into a common feature space, illustrated in Figure **??**. Modern research is now focused on combining Hyperalignment with large-scale pre-training to enable zero-shot or few-shot transfer learning to new subjects.

A second major challenge lies in capturing Temporal Dynamics and Video Reconstruction. While static image reconstruction has achieved photorealism, reconstructing continuous visual experiences remains difficult because the BOLD signal is sluggish, acting as a low-pass filter with a delay of 4–6 seconds. Consequently, fast-moving visual changes are often blurred or lost entirely in the fMRI signal. Opportunities to resolve this involve multi-modal fusion, specifically integrating fMRI (which offers high spatial resolution) with Magnetoencephalography (MEG) or EEG (which offer high temporal resolution) to resolve millisecond-level dynamics. Recent work by Wang et al. (2022) [11] has begun addressing this by using continuous transformer models to decode "cinematic" experiences, yet real-time, frame-perfect reconstruction remains an unsolved inverse problem.

Finally, the field faces the Hallucination Problem. The reliance on strong generative priors, such as Stable Diffusion, creates a new class of error where the model generates plausible but incorrect details. For instance, a diffusion-based decoder might reconstruct a "red Ferrari" when the user viewed a "red Ford." While the image possesses high fidelity and semantic similarity, the specific instance is factually incorrect. In clinical contexts, such as communication for patients with locked-in syndrome, these hallucinations could lead to critical misunderstandings. As noted by Tang et al. (2023) [12] in the context of language decoding, the field requires new evaluation metrics that penalize plausibility when it deviates from ground truth. Future architectures must find a better balance between the generative prior (visual quality) and strict data fidelity (truthfulness to the brain signal).

References

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

[2] Miyawaki, Y., et al. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5), 915–929.

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

[4] Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1), e1006633.

[5] Seeliger, K., et al. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181, 775–785.

[6] Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. *CVPR*, 14453–14463.

[7] Chen, Z., et al. (2023). Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. *CVPR*, 22710–22720.

[8] Scotti, P. S., et al. (2023). Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. *NeurIPS*, 36.

[9] Scotti, P. S., et al. (2024). MindEye2: Shared-Subject Models Enable fMRI-to-Image with 1 Hour of Data. *arXiv preprint arXiv:2403.11207.*

[10] Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., ... & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2), 404–416.

[11] Wang, Z., Wu, J., Zhang, X., Li, C., & Guo, Y. (2022). Cinematic Mind: Cross-modal Language-Auditory-Visual Brain Decoding. *arXiv preprint arXiv:2207.09699.*

[12] Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5), 858–866.

[13] Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000: A public fMRI dataset of 5000 images. *Scientific Data*, 6(1), 1–19.

[14] Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1), 116–126.

[15] Urbanski, M., Coubard, O. A., & Bourlon, C. (2014). Visualizing the blind brain: brain imaging of visual field defects from early recovery to rehabilitation techniques. *Frontiers in Integrative Neuroscience*, 8(74).

[16] Nikroorezaei, F., & Saraf Esmaili, S. (2019). Application of Models based on Human Vision in Medical Image Processing: A Review Article. *International Journal of Image, Graphics and Signal Processing*, 11(12), 23–28.

[17] Rakhimberdina, Z., Jodelet, Q., Liu, X., & Murata, T. (2021). Natural Image Reconstruction From fMRI Using Deep Learning: A Survey. *Frontiers in Neuroscience*, 15, 795488.

[18] Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A Model of Representational Spaces in Human Cortex. *Cerebral Cortex*, 26(6), 2919–2934.