CS292F final report: image reconstruction from fMRI

Yuchen Hou
Department of Computer Science
University of California, Santa Barbara
yuchenhou@ucsb.edu

Setareh Najafi Khoshnoo
Department of Computer Science
University of California, Santa Barbara
setareh@ucsb.edu

1 Introduction and related works

With the recent advancements in machine learning and neuroscience, the reconstruction of images from brain activities has become one of the most popular brain-decoding tasks [1]. The brain-to-image reconstruction could provide valuable insights into neural coding mechanisms underlying visual perception, working memory, brain states, and mental imagery. However, this task remains challenging due to the complex nature of training generative models, the sparsity of the training data availability, and the low signal-to-noise ratio in recordings [2].

Early deep-learning approaches in reconstructing images from fMRI (functional Magnetic Resonance Imaging, a non-invasive neural recording technique) focused on achieving pixel-level similarity with original images (e.g., [3]). However, these early reconstructions often lacked semantic clarity, resulting in blurry and unrecognizable outputs [3].

Recent works have leveraged advanced deep generative models like Generative Adversarial Networks (GAN), Variational Autoencoders (VAE), and Latent Diffusion Models (LDM) to produce high-resolution images with rich semantic information. For example, Takagi & Nishimoto utilized two linear regressions to learn mappings from fMRI data to latent image feature z derived from VQ-VAE image encoder and text feature c derived from CLIP text encoder within a pre-trained Stable Diffusion [2]. Without fine-tuning the Stable Diffusion model, they were able to achieve state-of-the-art reconstruction of complex images [2]. Lu et al. [4] expanded upon this work by introducing a third linear mapping from fMRI to the CLIP image encoder's latent space and continuously optimizing the learned latent features through back-propagation. Compared with its predecessor, their proposed two-stage training pipeline generated more accurate reconstructions in terms of shape, orientation, and position [4]. More recently, Ferrante et al. trained regularized linear regressors between brain activity and latent features extracted from the Generative Image-to-text Transformer to produce initial images, depth maps, and text captions, which were then refined using the ControlNet-augmented Stable Diffusion model to generate the final reconstructions [5].

While these studies made significant achievements, their approaches of relying on linear projections from fMRI features to pre-trained latent spaces may not fully capture the brain's complex neural coding. For example, the uneven distribution of cone photoreceptors in the retina [6] and the cortical magnification in the primary visual cortex [7] indicate the nonlinear nature of brain's visual information processing. This discrepancy between the linear fMRI mapping in [2, 4-5] and the nonlinear processing in the brain will make the fMRI-to-image generative models less representative of the actual brain function, potentially impairing the reconstruction performance.

To this date, limited studies have utilized nonlinear neural networks for task-relevant fMRI feature extraction for LDM. For example, [9] utilized 24 blocks of vision transformers with masked sparse coding to decode fMRI into embeddings and fed them into an LDM via cross-attention. However, this study lacks comprehensive comparisons across various nonlinear architectures. To fill this gap, we aimed to investigate how different nonlinear neural network architectures can enhance the performance of fMRI-to-image reconstruction.

Recently, graph-based networks have shown superior performance and explainability in decoding task-driven fMRI activities given their ability to represent the topological architecture and connectivity in human brain networks [10]. However, tasks involving fMRI-to-image decoding with graph-based architectures have predominantly concentrated on classification to enhance visual decoding accuracy (e.g., [11]). To this date, no study has incorporated Graph Convolutional Networks (GCN) in LDMs for image reconstruction. Motivated by the hypothesis that GCN's performance at mapping brain architectures could help decode fMRI data into more accurate feature representations, the current study also aimed to explore the potential of GCNs in image reconstruction from fMRI data.

This study has three major contributions:

1. We proposed four nonlinear neural network modules to map brain fMRI data into the corresponding image and text latent features in response to the visual stimuli.

2. We introduced a GCN-based architecture for our fMRI-to-image decoder, comparing it with two other architectures to evaluate its efficacy in predicting image latent features.

3. We adapted our architecture to two distinct datasets (NSD and THINGS-fMRI), demonstrating its potential for Brain-Computer Interface applications and establishing new benchmarks for future studies.

2 Methodology

2.1 Overview

In this section, we present our proposed architecture and training pipeline (Figure 1). In brief, we utilized a fMRI-to-text module to decode fMRI into CLIP text embeddings *c* and designed three fMRI-to-image modules to decode fMRI into VQ-VAE visual feature *z*. The learned latent features *c* and *z* were then used to condition the Stable Diffusion to reconstruct images.

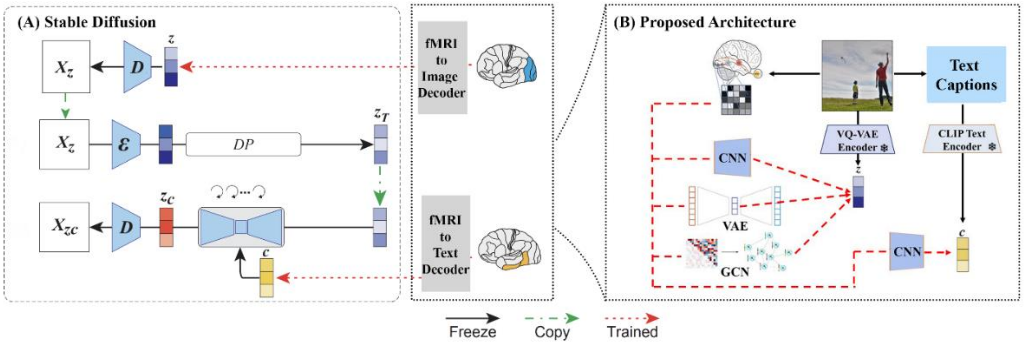

Figure 1. Overview of the proposed architecture. Image adapted from [2] and [4].

2.2 Datasets

We used two publicly available datasets: the Natural Scenes Dataset (NSD) [12] and the THINGS-fMRI dataset [13].

NSD dataset The NSD is one of the largest neuroimaging datasets using images sampled from Microsoft COCO dataset [12]. We focused on data collected from one subject ("subj01"). The training set contains 8,859 image stimuli across 24,980 fMRI trials, and the test set contains 982 image stimuli across 2,770 fMRI trials. In the test set, the fMRI responses were averaged across repetitions. The brain voxel Regions of Interest (ROIs) were identified using the provided surface-based ROI masks ("HCP_MMP1" and "streams"). Corresponding COCO captions for each image were extracted using the COCO ID of the stimuli.

THINGS-fMRI dataset Data from only one subject ("sub-01") was included in this study. The training set contains 8,640 trials, and the test set contains 1,200 trials with 12 repetitions per image. We averaged fMRI responses across repetitions in the test set. The descriptions collected from either WordNet, Google, or Wikipedia of the image categories [13] were considered as the text caption for images. The brain's ROI masks were directly derived from the provided files. Table 1 shows detailed voxel counts extracted from each ROI in both datasets.

Dataset	ROI	Num of Voxels
NSD	early visual cortex (V1-V4)	10631
	higher ventral stream	7604
THINGS-fMRI	early visual cortex (V1-hV4)	3198
	higher visual cortex (VO1, VO2, TO1, TO2, FFA, PPA, LOC)[1]	5183

Table 1: The ROI and the number of voxels in NSD and THINGS-fMRI used in this study.

2.3 Stable diffusion

To reconstruct images from brain fMRI activity, we utilized an LDM called Stable Diffusion (Figure 1A). During training, the forward diffusion process gradually corrupts the structure of the input image by adding Gaussian noise ϵ for each time step *t.* The reverse denoising process, parameterized by Gaussian transition, learns a noise model at time t to reverse the added noise from the latent image. This diffusion-denoising procedure can be conditioned by initial latent image features z and latent features c (decoded from various modalities via cross-attention) [17]. In our experiment, z is derived from mapping fMRI to the latent ground-truth image features and undergoes a forward diffusion process for 40 steps to obtain a noise-added initial feature z_T, and c is derived from mapping fMRI to the latent ground-truth text features. The denoising U-net structure then integrates z_T and c, which is fed into VQ-VAE decoder to reconstruct the results. Overall, the Stable Diffusion model aims to minimize the following training objective:

$$L = E_{z_t,\ y,\ \epsilon\sim N(0,1),\ t}[||\epsilon - \epsilon_\theta(z_t, t, c))||_2^2]\ [2]$$

Throughout the entire experiment, the weights of the Stable Diffusion model were frozen, and the reconstructed images were of size 512×512 pixels.

2.4 fMRI-to-text module

We first passed each image's captions into pre-trained Stable Diffusion's CLIP text encoder to obtain the ground-truth latent text features. If there were multiple captions for an image, the corresponding latent features would be averaged across captions. We then reduced each trial's latent text feature from 77×768 (tokens × embedding dimensions) to 15×768 as indicated by [4] that 15 tokens were enough to cover the typical lengths of COCO captions.

We then implemented a CNN-based module to nonlinearly map fMRI data to text embeddings. We utilized fMRI data from the higher ventral stream (NSD) or higher visual cortex (THINGS-fMRI) as the input as these regions are more relevant to the semantics of visual scenes [2]. Inspired by [8], our CNN-based module contains a Conv1D layer, three residual blocks, another Conv1D layer, and three fully connected layers (1024 and 2048 hidden sizes for intermediate layers). Each residual block has three Conv1D layers (kernel size 7, padding 3), and any other Conv layers' kernel size is 1. Except for the last layer, each layer is followed by a ReLU nonlinearity.

2.5 fMRI-to-image modules

To decode brain fMRI into the latent image features, the ground-truth latent image features were obtained by inputting the original image stimuli into the pre-trained Stable Diffusion VQ-VAE's image encoder and extracting latent space features (4×40×40 dimensions) of each image. The fMRI data from lower visual areas (V1-V4 in NSD and V1-hV4 in THINGS-fMRI) were used for this module as these cortical regions are more relevant to low-level image features such as edges, contrast, orientations, or colors [2-4]. We constructed and compared three different architectures for mapping fMRI data to latent image spaces.

2.5.1 CNN

Adapted from [8], our CNN-based architecture is a modified version of our fMRI-to-text decoder. It includes one Conv1D layer, two residual blocks, another Conv1D layer, and three fully connected

[1] Other combinations had been attempted. It was observed that different cortical combinations did not yield substantial differences in our proposed neural networks but did affect [2]'s linear models to a certain extent.

layers (512 hidden size for intermediate layers). Each residual block has two Conv1D layers (kernel size 11, padding 5), and all other Conv layers' kernel size is set to 1. Except for the last layer, each layer in the architecture is followed by a ReLU nonlinearity.

2.5.2 VAE

The encoder of our VAE consists of two fully connected layers separated by a LeakyReLU nonlinearity [1]. It encodes flattened fMRI data and outputs the mean and logarithm of the variance of the latent image distribution. The decoder has three blocks with skip connections, each containing a fully connected layer, a BatchNorm1d layer, and a LeakyReLU layer.

2.5.3 GCN

GCN overview GCNs update node representations by aggregating and transforming features from neighboring nodes. This process involves the Laplacian matrix L of the graph, which encodes the structure of the graph. For a graph *G* with *N* nodes, the normalized Laplacian is defined as

$$L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}},$$

where *A* is the adjacency matrix, I_N is the identity matrix, and *D* is the diagonal degree matrix with $D_{ii} = \sum_j A_{ij}$. The ChebNet, introduced by Defferrard et al. [14], utilizes the Chebyshev polynomials of the first kind to approximate the graph Laplacian's spectral filters. It uses a truncated expansion of the Chebyshev polynomials up to a certain order *K* to create localized graph convolution filters. The convolution operation in ChebNet is defined as

$$g_\theta \star x = \sum_{k=0}^{K} \theta\ _k T_k(\tilde{L}) x$$

where θ represents the trainable parameters, T_k denotes the Chebyshev polynomial of order *k*, $\tilde{L} = 2L/\lambda_{max} - I_N$is the scaled Laplacian, and λ_{max} is the largest eigenvalue of L.

fMRI graph representation Following [15], we transformed the raw fMRI into graph representations. The graph nodes were the ROIs provided by the NSD or THINGS-fMRI dataset. Data from V1 and V2 were treated as two separate nodes, and data from V3 and V4 were combined into another node. The graph node's features were the corresponding ROI's normalized voxel activity. Each node had *k* features, and if the number of voxels was smaller than *k*, it would be zero-padded to *k*. The construction of graph edges had two steps. First, we averaged voxels' activities across all trials within each node and computed the functional connectivity across nodes using Pearson's correlation coefficient to obtain an adjacency matrix *M* of shape $N_{nodes} \times N_{nodes}$. We then thresholded the coefficients (edge attributes) to be 0.5, set the diagonal coefficients in *M* to zero to remove self-loops [15]. We also set M_{ij} to be zero for all $i, j \in N_{nodes}$ if $i > j$. This unidirectional design of the fMRI graph representation was motivated by the predominant bottom-up stimulus-driven nature of lower visual cortex activities [16].

Modular architecture Our proposed model consists of two ChebConv layers with 512 neurons (filter size is 4 for the first layer and 3 for the second) and with a BatchNorm layer and a ReLU nonlinearity in between. The global mean pooling layer is applied to get the vectorized latent representation, which is then passed into a Conv1D layer and two residual blocks. Each residual block consists of one Conv1D layer (kernel size 3, padding 1) and a LeakyReLU. Then, the latent feature is downsampled to 1 channel, flattened, and processed by two fully connected layers (1024 hidden size) to get the final latent image representation.

2.6 Training pipeline

Four modules (one fMRI-to-text and three fMRI-to-image, Figure 1B) were separately trained. Model parameters were optimized with Adam for 100 epochs to minimize the mean square error

(MSE) loss between the predicted latent feature y_{pred} and the corresponding ground-truth latent feature y_{actual} across all *N* image-fMRI pairs using

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{actual_i} - y_{pred_i})^2$$

In VAE, the loss during training includes a fMRI reconstruction loss and an image latent feature prediction loss. An early stopping with a patience of 5 epochs was used. Module-specific hyperparameters can be found in Table 2. All models were implemented in PyTorch and trained on an NVIDIA RTX 3090 with 24GB of memory. The code is available on GitHub: github.com/subawocit/CS292F_Project.

	Module			
Parameter	fMRI-to-Text: CNN	fMRI-to-Image: CNN	fMRI-to-Image: VAE	fMRI-to-Image: GCN
Learning Rate (NSD)	0.0001	0.0005	0.0005	0.001
Learning Rate (THINGS-fMRI)	0.001	0.0005	0.0005	0.0015
Batch Size	16	256	256	64

Table 2. Module-specific hyperparameters.

2.7 Evaluation

We computed the MSE scores between the predicted and the ground-truth latent features for each proposed module and compared them with the MSEs obtained by reproducing [2]'s predictions. Since our fMRI-to-image module has three variations, we selected the one with the highest performance on the test set as our final fMRI-to-image decoder. The predicted latent image and text features were then fed into the frozen Stable Diffusion model for image reconstructions. We used the Structural Similarity Index (SSIM), Per-pixel Correlation Coefficient (PCC), Learned Perceptual Image Patch Similarity (LPIPS) with an AlexNet backbone [9], and the average correlation distance using EfficientNetV2-S to quantitatively compare the performance of our proposed modules with [2]. To reduce biases, we generated 5 samples for each set of learned latent features, and the quantitative scores for each image stimulus were averaged across 5 samples.

3 Experiments and results

3.1 Feature decoding experiments

We trained our four architectures alongside [2]'s two L2-regularized linear regression models to map the fMRI activities to either CLIP latent text feature space or VQ-VAE latent image feature space (see Methods). We then computed the MSEs between the predicted values and the ground-truth features extracted from Stable Diffusion components. As shown in Table 3, for both datasets, our proposed CNN-based fMRI-to-text module and all of the fMRI-to-image modules outperformed [2]'s linear models in terms of MSE scores, highlighting the ability of our approach to adapt to variations across distinct datasets. Among the three proposed fMRI-to-image modules, the GCN-based architecture consistently yielded the best performance in terms of the MSE score[2].

3.2 Image reconstruction results

We then focused on image reconstruction using the latent text features predicted by our fMRI-to-text decoder and the latent image features from our best-performing fMRI-to-image decoder. These learned features were used to condition the Stable Diffusion model, which was kept frozen for this process. To compare with the baseline fairly, we also reproduced [2]'s results. As Figure 2 indicates, our methods yielded better reconstruction results qualitatively in the NSD dataset. The images

[2] The difference between the CNN- and GCN-based models was not large. A researcher at NeurIPS speculated that it was because of the residual connections within the CNN-based module that provided information about brain connectivity.

reconstructed by our approach showed semantically richer content and were structurally closer to their ground-truth counterparts compared to the baseline. Quantitatively, the proposed method again outperformed the baseline across both low- and high-level image features. To further demonstrate our approach's robustness, we tested the same architecture using the THINGS-fMRI dataset. While the results from this dataset were not as pronounced as those from the NSD dataset, our proposed method still showed superior performance over [2].

		Dataset	
Module	Method	NSD	THINGS-fMRI
fMRI-to-Text	Linear [2]	0.7283	1.1444
fMRI-to-Text	CNN (ours)	**0.2656**	**0.7941**
fMRI-to-Image	Linear [2]	0.7535	0.7432
fMRI-to-Image	CNN (ours)	0.5579	0.6920
fMRI-to-Image	VAE (ours)	0.5665	0.7260
fMRI-to-Image	GCN (ours)	**0.5540**	**0.6873**

Table 3. MSE results for fMRI-to-text and fMRI-to-image modules comparing our proposed methods with [2]. Lower scores represent better performance.

3.2 Image reconstruction results

We then focused on image reconstruction using the latent text features predicted by our fMRI-to-text decoder and the latent image features from our best-performing fMRI-to-image decoder. These learned features were used to condition the Stable Diffusion model, which was kept frozen for this process. To compare with the baseline fairly, we also reproduced [2]'s results. As Figure 2 indicates, our methods yielded better reconstruction results qualitatively in the NSD dataset. The images reconstructed by our approach showed semantically richer content and were structurally closer to their ground-truth counterparts compared to the baseline. Quantitatively, the proposed method again outperformed the baseline across both low- and high-level image features. To further demonstrate our approach's robustness, we tested the same architecture using the THINGS-fMRI dataset. While the results from this dataset were not as pronounced as those from the NSD dataset, our proposed method still showed superior performance over [2].

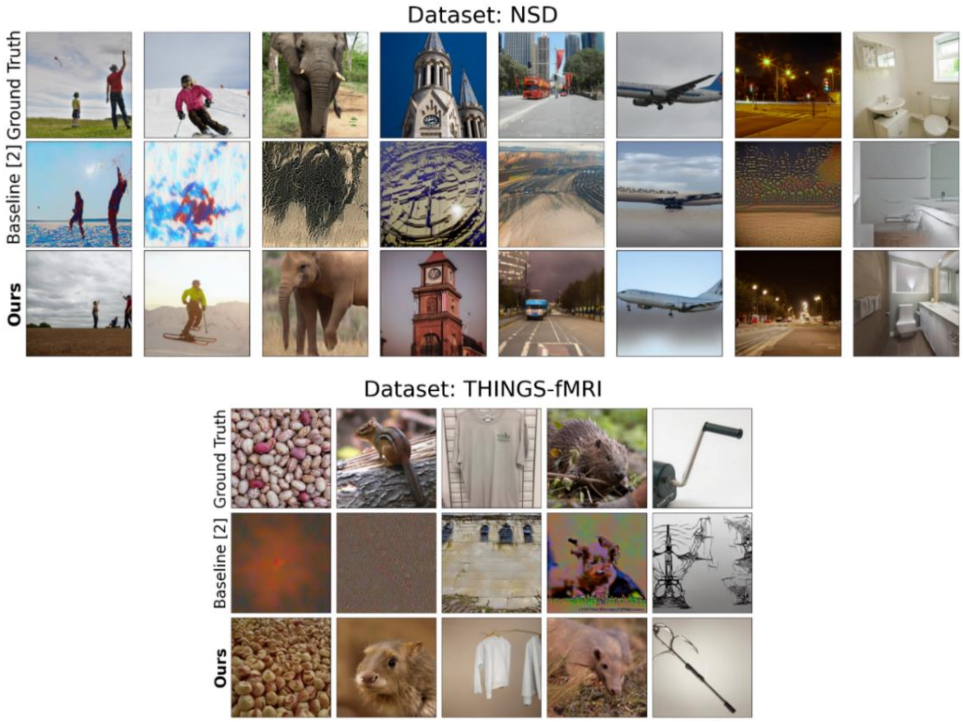

Figure 2. Reconstructed results.

Dataset	Method	PCC↑	SSIM↑	LPIPS (AlexNet) ↓	EfficientNet ↓
NSD	Linear [2]	0.2659	0.2141	0.7695	0.2791
	Ours	**0.3246**	**0.3528**	**0.7329**	**0.2734**
THINGS-fMRI	Linear [2]	0.1293	0.2407	0.8364	0.2556
	Ours	**0.2013**	**0.3993**	**0.8148**	**0.2454**

Table 4. Quantitative comparisons with the baseline.

4 Conclusions and future works

In this study, we proposed four fMRI decoding modules designed to map the brain fMRI data into the latent features derived from the CLIP text encoder or the VQ-VAE image encoder for image reconstruction. Our nonlinear architectures consistently outperformed the linear approach proposed by [2], achieving a higher structural similarity and richer semantic fidelity in image reconstructions across both NSD and THINGS-fMRI datasets. Our work is the first study using the THINGS dataset with fMRI recordings for image reconstruction, thereby setting a benchmark for future studies.

It is worth noting that, among the three proposed fMRI-to-image architectures, our GCN-based model surpassed the other two models in terms of decoding the brain activity into the latent image features. This observation aligns with the strengths of graph neural networks in capturing the complex topological and functional activities across brain regions. Additionally, we showed our methods' ability to adapt to another dataset by using the architectures designed for NSD dataset to fit the THINGS-fMRI dataset with minimal modifications. All these results highlight the importance of building brain-like models in brain decoding tasks.

Our study has limitations. First, we observed that the generation performance discrepancy between THINGS-fMRI and NSD is large, and future work should still look for better architectures designed across different datasets. Second, our pipeline requires extracting latent features from the stimuli before training the brain decoders. It is possible that a cross-subject, end-to-end, and multimodal joint training approach would have more control over the generation procedure and therefore generate better results. Third, we flattened the fMRI activity into a 1D vector for each brain ROI. Since the spatial topology of the 2D cortical areas within each ROI is better at capturing the brain connectivity pattern [19], future work should utilize this topological structure for brain decoding. Fourth, this work only normalized the fMRI data for preprocessing. However, the brain encodes information in a more sophisticated manner: for example, only <1% of neurons in V1 are strongly activating upon visual stimuli [20]. Therefore, it is important to explore different neural encoding/dimensional reduction strategies. Fifth, although our architecture designs are aligned with the brain functionality, its neuro-interpretability is unknown. In the future, one can use our brain-inspired modules with the NSD dataset to fine-tune the Stable Diffusion model and investigate the neurological implications of visual scene reconstruction and the correlation between model architecture and the brain [18].

In conclusion, this study introduces four brain-to-stimuli decoding methods and shows the superior capability of nonlinear brain-inspired architectures in reconstructing images from fMRI data, providing potential insights into visual reconstructions for Brain-Computer Interfaces.

References

[1] Y. Liu, Y. Ma, W. Zhou, G. Zhu, and N. Zheng, “BrainCLIP: Bridging Brain and Visual-Linguistic Representation Via CLIP for Generic Natural Visual Stimulus Decoding,” *arXiv*, May 2023.

[2] Y. Takagi and S. Nishimoto, “High-resolution image reconstruction with latent diffusion models from human brain activity,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi:10.1109/cvpr52729.2023.01389

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

[4] Y. Lu, C. Du, Q. Zhou, D. Wang, and H. He, “Minddiffuser: Controlled image reconstruction from human brain activity with Semantic and structural diffusion,” *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. doi:10.1145/3581783.3613832

[5] M. Ferrante, F. Ozcelik, T. Boccato, R. VanRullen, and N. Toschi, “Brain Captioning: Decoding human brain activity into images and text,” *arXiv preprint arXiv:2305.11560*, 2023.

[6] *“Anatomical distribution of rods and cones - neuroscience - NCBI bookshelf,”* https://www.ncbi.nlm.nih.gov/books/NBK10848/ (accessed Oct. 27, 2023).

[7] R. A. Cohen, “Cortical magnification,” *SpringerLink*, https://link.springer.com/referenceworkentry/10.1007/978-0-387-79948-3_1355 (accessed Oct. 27, 2023).

[8] S. Lin, T. Sprague, and A. K Singh, “Mind Reader: Reconstructing complex images from brain activities,” *NeurIPS*, 35:29624–29636, 2022.

[9] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. Zhou, “Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding,” *CVPR*, 2023

[10] X. Li *et al.*, “Braingnn: Interpretable brain graph neural network for fmri analysis,” *Medical Image Analysis*, vol. 74, p. 102233, 2021. doi:10.1016/j.media.2021.102233

[11] L. Meng and K. Ge, “Decoding visual fmri stimuli from human brain based on graph convolutional neural network,” *Brain Sciences*, vol. 12, no. 10, p. 1394, 2022. doi:10.3390/brainsci12101394

[12] E. J Allen *et al.*, “A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence,” *Nature neuroscience*, 25(1):116–126, 2022.

[13] M. N. Hebart *et al.*, “Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images,” *PLOS ONE*, vol. 14, no. 10, 2019. doi:10.1371/journal.pone.0223792

[14] M. Defferrard, X. Bresson and P. Van der Gheynst, "Convolutional neural networks on graphs with fast localized spectral filtering", *Proc. NIPS*, 2016.

[15] M. Saeidi *et al.*, “Decoding task-based fmri data with graph neural networks, considering individual differences,” *Brain Sciences*, vol. 12, no. 8, p. 1094, 2022. doi:10.3390/brainsci12081094

[16] S. McMains and S. Kastner, “Interactions of top-down and bottom-up mechanisms in human visual cortex,” *The Journal of Neuroscience*, vol. 31, no. 2, pp. 587–597, 2011. doi:10.1523/jneurosci.3766-10.2011

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with Latent Diffusion Models,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2022. doi:10.1109/cvpr52688.2022.01042

[18] C. Conwell, J. S. Prince, K. N. Kay, G. A. Alvarez, and T. Konkle, *What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?,* 2022. doi:10.1101/2022.03.28.485868

[19] Z. Gu, K. Jamison, A. Kuceyeski, and M. Sabuncu, “Decoding natural image stimuli from fmri data with a surface-based convolutional network,” In *MIDL*, 2023.

[20] T. Li, Z. Wen , Y. Li , and T. S. Lee, “Emergence of Shape Bias in Convolutional Neural Networks through Activation Sparsity,” *NeurIPS*, 2023.