Deep Neural Network Decoding of Image Surrounds from V1 Signal Recordings

Technical Report

Hadrien Helfgott

Carnegie Mellon University

15 689: Independent Study in the Computer Sciences

Under the supervision of:

Pr. Tai-Sing Lee

April 2024

S. No	Contents	Page(s)
1	Introduction	3
2	Dataset	7
3	Methods	10
4	Results	16
5	Discussion	19
6	References	21

1. Introduction

At the beginnings of Machine Learning as we know it, most of neural networks' appeal to the general public was their seemed ability to replicate the inner workings of the human brain. Since then, a decade of innovation has shed light on a more complex reality: the models that we use and develop have more to do with a trial-and-error based search for performance than with replicating biological neurons. In fact, such a task currently seems impossible: most recent attempts are reproducing full human brains with neural networks such as the Human Brain Project quickly hit a wall of resources, complexity and knowledge gaps (Frégnac, 2023). We are still very ignorant of the inner mechanisms of our brains, and still mostly unable to understand how neurons and electrical signals give rise to our perception or cognitive abilities.

In this context, another question arises. After using high-level ideas from the brain's structure to construct artificial neural networks, can we use these programs to understand our own brains with more depth than we ever have? One of most prominent fields of research that could benefit from this approach is visual reconstruction. Visual reconstruction seeks to use brain signals to re-create an image that a subject was seeing at the time where their signals were recorded. While useful for Brain-Computer Interfaces, successful results in this field also have a broader scientific interest as they would teach us meaningful insights about human perception as a whole.

Historically, Visual Reconstruction has been done in many different ways, most of them focusing on broad, low-precision visual features, and attempting to classify rather than decode stimuli from scratch. Kamitani & Tong (2005) and Naselaris et al. (2011) used statistical mathematical methods to decode which orientation a subject was seeing. Stokes et al. used in 2009 an early linear regression model to decode whether subjects were seeing X's or O's on a board. As most papers on this topic, these studies used fMRI, meaning they were often limited

to images instead of time series due to fMRI's poor temporal resolution (Nishimoto et al., 2011). This also means that most analyses relied on fMRI's high spatial accuracy to create retinotopic maps, and then classify or decode easily identifiable visual patterns (Thirion et al., 2006; Miyawaki et al., 2008). In other words, early decoding works focused on identifying which parts of an image were paid attention to, and deducing the image from this information, rather than trying to directly decode the image's content. It should be noted that such an approach has its limits, given the high overlap in fMRI activation of different patterns, even with objects known to be processed with very high discriminability such as faces versus objects (Haxby et al., 2001). A first important step in more precise content reconstruction, such as natural scenes, was the use of Gabor features (Kay et al., 2008). Gabor features are 2D very simple visual primitives, hypothesized since the 1980's to form the base blocks of visual perception (Daugman, 1985) due to their ability to represent complex 2D images when combined (Lee, 1996). Since then, more recent works have harnessed the power of generative AI models to best use these Gabor features in neural modeling works (Cui et al., 2021).

A second major development in the visual reconstruction has been the rise of generative AI models over the last decade, particularly generative adversarial networks (GAN's) and variational auto-encoders (VAE's) (Du et al., 2022). VAE's are deep auto-encoders that train to minimize the KL divergence between their latent space distribution and the normal, resulting in stable training and a continuous latent space. They have been extensively used for fMRI brain decoding , for example to decode handwritten digits (Du et al., 2019) or natural sceneries (Shen et al., 2019) with mixed results: attempts at decoding complex visual stimuli often result in an undefined color mix where shapes of the original objects are identifiable but very imprecise. On the other hand, GAN's use a generator and a decoder to progressively train a model to create, and discriminate, more and more realistic stimuli. Applied to brain signals, it can yield similar and sometimes better results than VAE in natural scene reconstruction (Huang et al., 2020;

Huang et al., 2021). The results are often less blurry, more realistic-looking and more defined than VAE (Seeliger, 2018), but they suffer from other major shortcomings including a tendency to hallucination (Seeliger, 2018; Huang et al., 2020) and overall variety issues (Hayasi & Kawata, 2018) .

A major common factor underlies all the papers discussed above: they all use fMRI. While highly precise, fMRI is also slow and costly. It is particularly misfit for dynamic analysis and real-time computations using brain signals (Benchetrit, Banville, King, 2023). In a context where neural decoding's main motivation is often real-time Brain-Computer Interfaces (Du et al., 2023), it is crucial to develop a better understanding of how to decode neural signals with faster, albeit more noisy, neural recording devices. Many recent works have used EEG signals in conjunction with ML models for neural decoding, SVM being the most used and most accurate machine learning model (Saeidi et al., 2021) and CNN coming second in terms of precision. In 2020, Zheng et al. used EEG to train a GAN and successfully classify images. However, the majority of EEG-based neural encoding has little to do with visual imagery as the most common explored topics are emotional recognition, mental workload, and motor imagery (Zheng et al., 2020).

Another technique, two-photon calcium imaging (Stosiek et al., 2003; Grienberger et al., 2022), has been used to obtain high-temporal, high-spatial resolution of localized patch of the brain. It is invasive, but its small scale allows for recording in freely moving and non-anesthetized animals. Two-photon calcium imaging's asset is its precision when analyzing real-time activity of restricted neural populations, although it can also be used effectively for single-cell recordings (Helmchen, 2009). The latter was used to collect the dataset used in the present report. Although non comparable in precision with fMRI, single-cell two-photon calcium imaging proved to be able to identify some very basic shapes in images (Garasto, Bharath & Schultz, 2018) and achieves good reconstruction performance when given artificial simple

images (Zhang et al., 2022). However, full natural images' reconstructions are still unrecognizable, likely due to the overall noisiness of the single-cell recordings (Malik et al., 2011). The current report attempts at reconstructing low-resolution images of natural scenes, using single-cell, single-timepoint recordings from 330 neurons.

When doing visual decoding using neural signals, an important question is where to look. Early visual cortices are widely considered to encode low-level features over small receptive fields, while late visual cortices encode full-field and high-level concepts. However, it has been hypothesized that early visual cortices may have an important role in scene recognition without any explicit object labeling typically obtained from higher cortices (Groen, Silson & Baker, 2017) and that early cortices may show a preference for natural images due to their sensitivity to these image's statistical regularities (Coggan et al., 2017). In addition, it has been shown that neurons in V1 are modulated by their receptive field (RF) surround (Nurminen & Angelucci, 2014) in animals as diverse as cats (Walker, Ohzawa & Freeman, 2000) and monkeys (Contributions of low- and high-level properties to neural processing of visual scenes in the human brain et al., 2009). Most of these modulations have been characterized as suppressive (Walker, Ohzawa & Freeman, 2000; Cavanaugh, Bair & Movshon, 2002; DeAngelis, Freeman & Ohzawa, 1994) and sensitive to orientation (Sengpiel, Sen & Blakemore, 1997; Shushruth et al., 2012) . This indicates that the brain signal recordings of single neurons in V1 could give us information not only on these neuron's own receptive field contents, but also on its surroundings through information from neighboring neurons. In this context, the present report tries to apply this to analysis of decoding performance.

If V1 neurons only hold information about their own receptive field, there should be a steep drop in reconstruction performance as soon as we try to re-create images broader than what's contained in these neuron's receptive field. If not, this could be a sign that neurons with a localized field can hold information about the broader picture that is being seen, whether it is

due to feedback connections from higher cortices (Chen et al., 2021) neighboring modulation, or the overall predictability of the peripheral field of natural scenes. This could prove particularly relevant when analyzing natural scenes, as it has been shown that peripheral vision is particularly useful, perhaps even more than central, for natural image recognition (Wang & Cottrell, 2017). In the present report, we therefore seek to verify the two following hypotheses:

- It is possible to extract visual information and achieve significant reconstruction performance from 2-photon calcium imaging data.
- This reconstruction performance will at least be partially maintained when trying to reconstruct images broader than the observed neurons' visual field.

2. Dataset

2.a) Data Acquisition

Two-photon imaging with GCaMP5 was used to measure the responses of 1689 neurons across 6 different sites of 3 macaque monkeys of 30k to 50k images. On each site, roughly 300 cells were tracked across 5 days during a fixation task. The images were presented in sequence for a duration of 1 second each, with 1 second of gray screen in between. The present report only focuses on a single site corresponding to a square millimeter of one monkey's V1 cortex. It contains exactly 330 neuron recordings for 50700 images. Neurons were tracked and registered across days based on Pearson correlation of neural responses to 200 fingerprint images, which were presented every day.

Aside from this 50k set, which will be referred to as « Training Set » for the rest of the report, the monkeys' responses to a 1000-image « Validation Set » were recorded using the same protocol across all 5 days. These images were all presented a total of 10 times each.

By conducting different analyses on the dataset, it was possible to highlight some of its key properties. Although informative in the context of this experiment, this analysis could also be useful in future 2-photon-calcium images studies as an understanding of the very nature of such datasets.

2.b) Data Analysis

There is very little redundancy in neuron responses

A PCA reduction of the Training Dataset revealed that the cumulated variance across principal components was very close to a linear curve (AUC = 0.608). This indicates very little redundancy in the brain signals, which tends to show that neurons share little if no correlation. This apparent impossibility to observe reliable repeating, redundant patterns in the data may be an indicator of multiple things. This may simply be due to the selection of neurons, that would be focused on distinct and unrelated features of the inputs. This could also be due at least partially to random noise inherent to the 2-calcium photon imaging procedure (Malik et al., 2011)

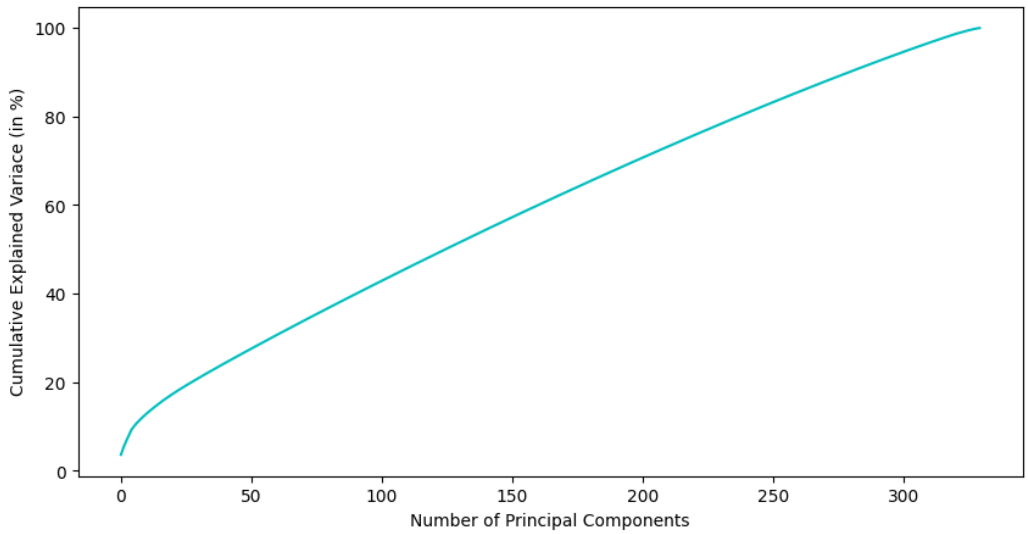

Figure 1. Cumulative Variance explained by principal components of the Training Dataset

Neuron responses are highly noisy

A 1-NN classifier was run on the Validation Set based on response-to-response Euclidian distance. For each response in the validation set, there are 9 other responses to the same image. In a non-noisy context assumption, responses to the same image should be highly similar, leading to a high 1-NN accuracy: each response's closest neighbor would be one of the 9 other responses to the same image. The actual observed classifier accuracy was 9.9 %. Although clearly better than chance (0.1 %), this result reveals high same-response, across-trial variability within the dataset.

A second classifier was built using the following process: for each image in the Validation Set, its first trial's response was compared against the mean of the responses to the other nine trials of each image. Then, the set of nine responses whose mean was the closest to the first example was selected as corresponding to the right image. The accuracy of the classifier was 17.6 %. This is still much higher than chance (0.1 %) but means that only a fifth of the images can be reliably assessed by observing their distance with 9 other trials of the *same* image in a 1000-image dataset.

There is no such thing as a group of good or bad neurons

Discriminability of neurons can be assessed by the standard deviation of their mean response to different images. The higher the standard deviation, the better individual neurons can discriminate between different images. On the other hand, noisiness can be assessed by the mean of within-image response standard deviations. If a given neuron consistently shows high standard deviation between the trials of a same image, it can be considered noisier than a more stable neuron. A correlation of 0.3126 ($p = 6.771 * 10e-9$) was observed between these two measures, meaning that neurons that are more discriminant across-image are typically also noisier within-image. In this context, attempting to pick a group of highly across-image

discriminant neurons to drive further analysis while discarding within-image noisy neurons is unlikely to be successful.

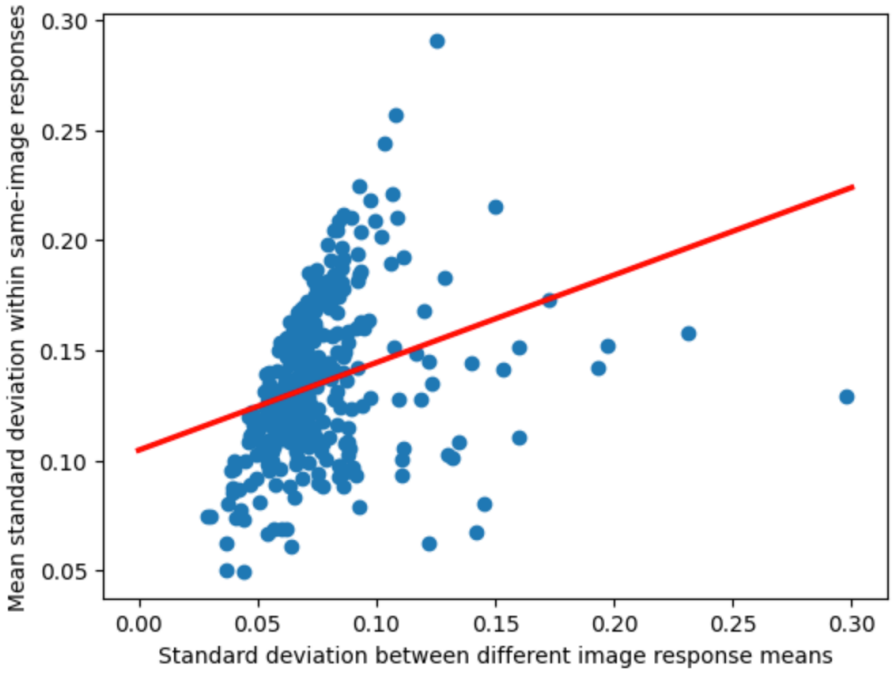

Figure 2. Correlation between within-image response variability and across-image response variability within the Training Dataset

3. Methods

3.a) Inputs

In order to analyze the brain responses, two different types of inputs were put together. The first one is the raw response, a 330-long vector with each value corresponding to the response of a single neuron to an image. We will refer to this type of input as a *response vector*. The second one is a 2d map representing the responses with respect to their actual position on the square of V1 cortex that was recorded. A 35 x 35 figure representing this 1 mm square was computed, and each single response was precisely mapped to the position of its corresponding

neuron on the square. When two neurons overlapped on the square, only the maximum response was kept. We will refer to type of input as a *response map*. While the response vector corresponds to the raw signal, the response map allows us to analyze responses with more accuracy with respect to their spatial component, allowing our network to capture potential noise dependencies in closer neurons.

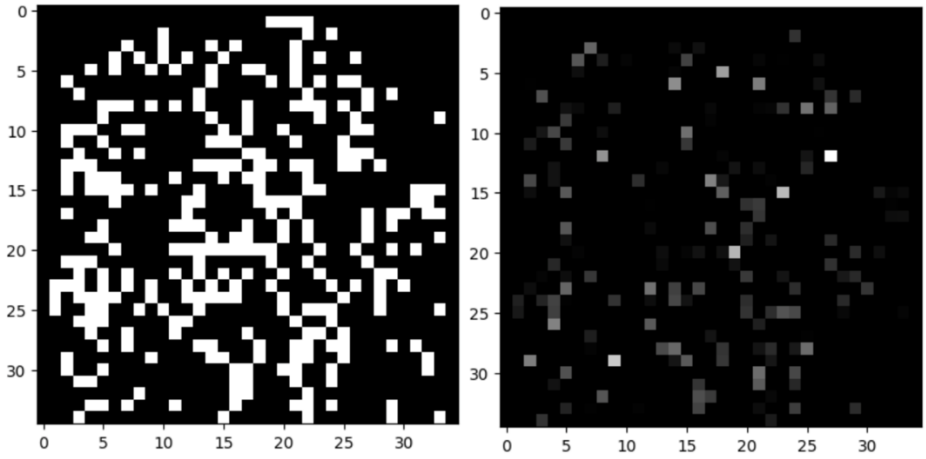

Figure 3. Left: Location of all neurons (white) on the 35 x *35 response map corresponding to the observed square of V1. Right: Actual response map of brain signals. Whiter tones correspond to higher intensity.*

3.b) Mapping Images to Latent Encodings using a VAE

In order to make the decoding easier, we decided to map the images to a 32-dimension latent space using a Variational Auto-Encoder. This way, instead of predicting full images from raw brain responses, we would predict a 32-long encoding vector which could then be used to reconstruct the image using the decoder half of that same VAE. In this report's experiments, images from 20 x 20 to 100 x 100 pixels were decoded. A separate VAE was trained for each image size, but for consistency they all used the same architecture.

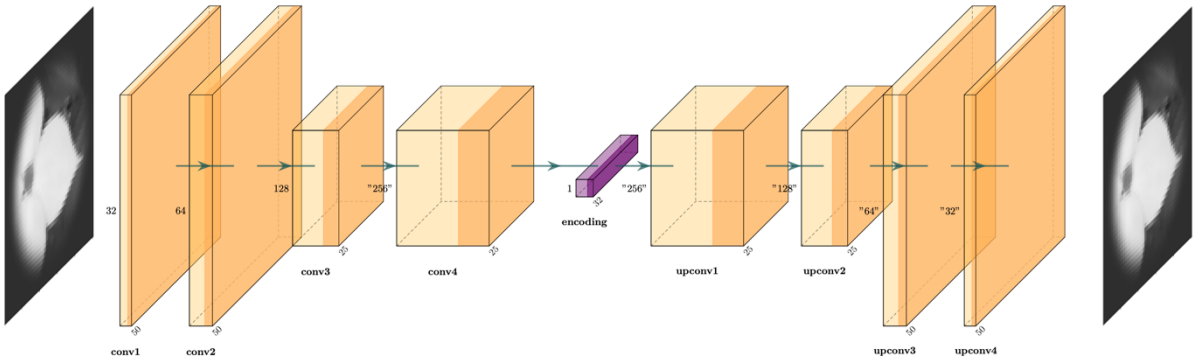

Figure 4. Visual representation of the VAE used to obtain latent space image encodings (purple)

	Layer	Output Shape	Kernel Shape	Stride	Padding
Image VAE	Conv2D	(50 x 50 x 32)	3 x 3 x 32	1	1
	Conv2D	(50 x 50 x 64)	3 x 3 x 64	1	2
	Conv2D	(25 x 25 x 128)	3 x 3 x 128	1	1
	Conv2D	(25 x 25 x 256)	3 x 3 x 256	1	1
	Linear	(1 x 1 x 32)			
	Sampling	(1 x 1 x 32)			
	UpConv2d	(25 x 25 x 128)	3 x 3 x 128	1	1
	UpConv2d	(50 x 50 x 64)	2 x 2 x 64	1	2
	UpConv2d	(50 x 50 x 32)	3 x 3 x 32	1	1
	UpConv2d	(50 x 50 x 1)	3 x 3 x 1	0	1

Table 1. Hyperparameters of the Image VAE. This example uses image size 50, but the VAE was adapted to each image size it was tested on. The internal convolution layers were set to half of the image size (25 here).

Each VAE was trained for exactly 5 epochs with batch-size 16, using PyTorch's learning rate scheduler. The training was done on 45 000 images of the Training Set, while the rest was used as a validation dataset.

3.c) Decoding Brain Responses using an Ensemble Network

In the first experiment, the goal was to obtain the maximum reconstruction performance for 100 x 100 full images. After mapping images to a latent space, the next step was to map

brain responses to this same latent space to allow image reconstruction. The following 3 parts were used:

- A CNN architecture followed by a gaussian sampling procedure (identical to the one used in VAE's) and 3 linear layers mapped the 2d response maps to a 32-long encoding vector. This structure is similar to a VAE Encoder followed by a MLP, the point being to use the spatial representation of neurons to map the response to a first lower-dimensional latent space before creating a linear-layers bridge between this space and the actual image encoding space.
- A classic Multi-Layer Perceptron mapped the 1d response vector to a 64-long encoding vector
- Both encoding from steps above, plus the actual raw response vector, were concatenated to form a 426-long vector which was then mapped to the image encoding previously computed by the Image VAE using a linear layer.

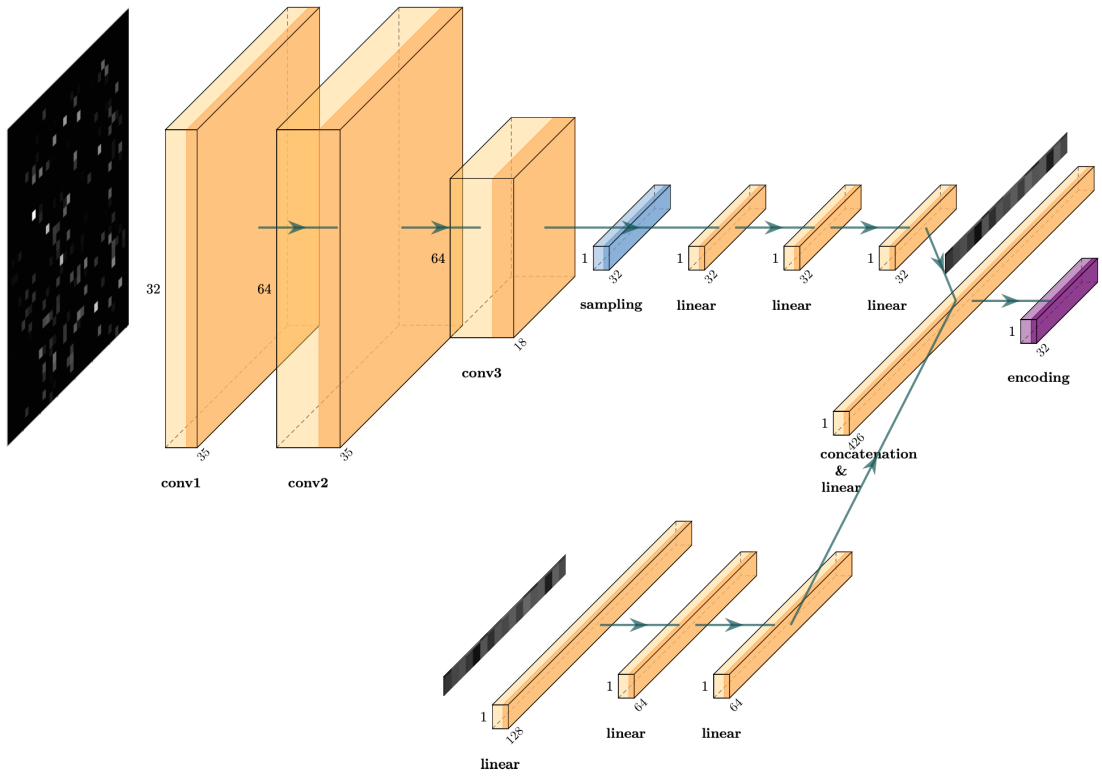

Figure 5. Visual representation of the ensemble network used to reconstruct previously computed latent space image encodings (purple)

	Layer	Output Shape	Kernel Shape	Stride	Padding
CNN Encoder	Conv2D	(35 x 35 x 32)	3 x 3 x 32	1	1
	Conv2D	(35 x 35 x 64)	3 x 3 x 64	1	2
	Conv2D	(18 x 18 x 64)	3 x 3 x 64	1	1
	Sampling	(1 x 1 x 32)			
	Linear	(1 x 1 x 32)			
	Linear	(1 x 1 x 32)			
	Linear	(1 x 1 x 32)			
MLP Encoder	Linear	(1 x 1 x 128)			
	Linear	(1 x 1 x 64)			
	Linear	(1 x 1 x 64)			
Ensemble Encoder	Concatenation	(1 x 1 x 426)			
	Linear	(1 x 1 x 32)			

Table 2. Hyperparameters of the ensemble network

The use of multiple paths (convolutional VAE encoder, MLP, and residual layer) was used to maximize reconstruction performance by crossing the insights gathered from different modalities and input types. The ensemble network was trained for exactly 5 epochs with batch-size 16, using PyTorch's learning rate scheduler. The training was done on 40 000 Images of the Training Set, while the rest were used as a validation set.

3.d) Varying Image Cropping

In order to conduct our second experiment, we created cropped versions of the full images. This was done by taking the 100 x 100 pixel original images and cropping their border so only the $n \times n$ center pixels remained. 3 center sizes were tried:

- 20 x 20 pixels, corresponding to the actual receptive field of the neurons

- 60 x 60 pixels, corresponding to the highest size that did not include the aperture (being constant on all images, this easily predictable aperture would have artificially augmented the decoding performance)
- 40 x 40 pixels, an intermediate size between the two above

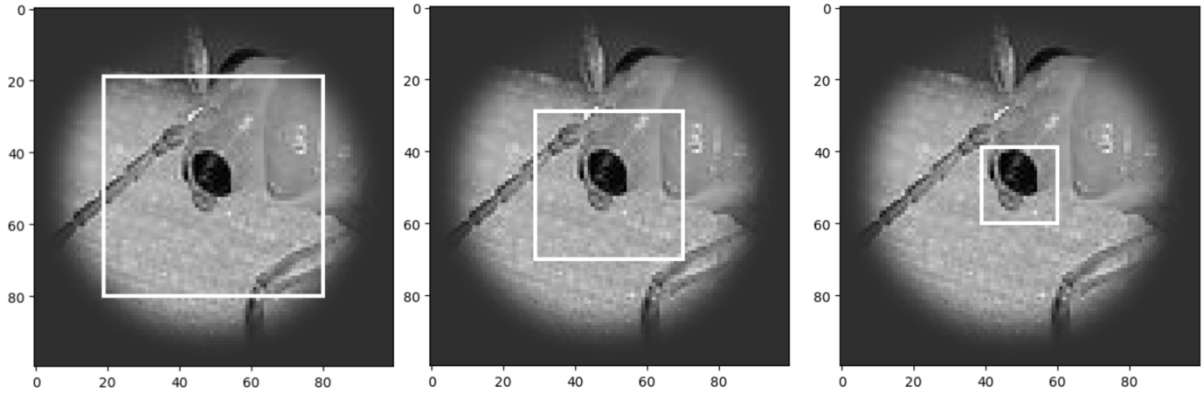

Figure 6. Cropping of Images to keep the 60-by-60, 40-by-40, and 20-by-20 centers

The different size experiments were conducted by first training a size-specific Image VAE (see 3.b) to compute the 32-dimensional encodings of all the images. Then, an ensemble encoder (see 3.c) was trained to map brain responses to these image encodings. The full process was repeated for each different center size.

3.e) Assessment of Reconstruction

In both experiments, reconstruction performance was assessed using the set containing the Validation images seen 10 times by the monkey. These images were not used in the training nor validation sets, they had therefore never been seen by either of our trained networks or used in their training in any way. For each of these images, a mean response was computed by averaging the 10 corresponding responses of the monkey. Then, these mean responses were input in the Ensemble Decoder to obtain a latent image encoding. The full image was

reconstructed by feeding these image encodings to the decoder half of our Image VAE. Performance was assessed using two metrics:

- Pixel-by-pixel correlation between the original image and the reconstructed one, averaged over all 1000 images of the Validation Set.
- Visual inspection of the reconstructions of the 32 first images of the set to get a qualitative insight of the model's performance

4. Results

4.a) Mapping Images to a latent space using a VAE

The first part of the decoding was mapping images to a 32-dimension latent space. The results of the Image VAE reconstruction for all sizes appears below:

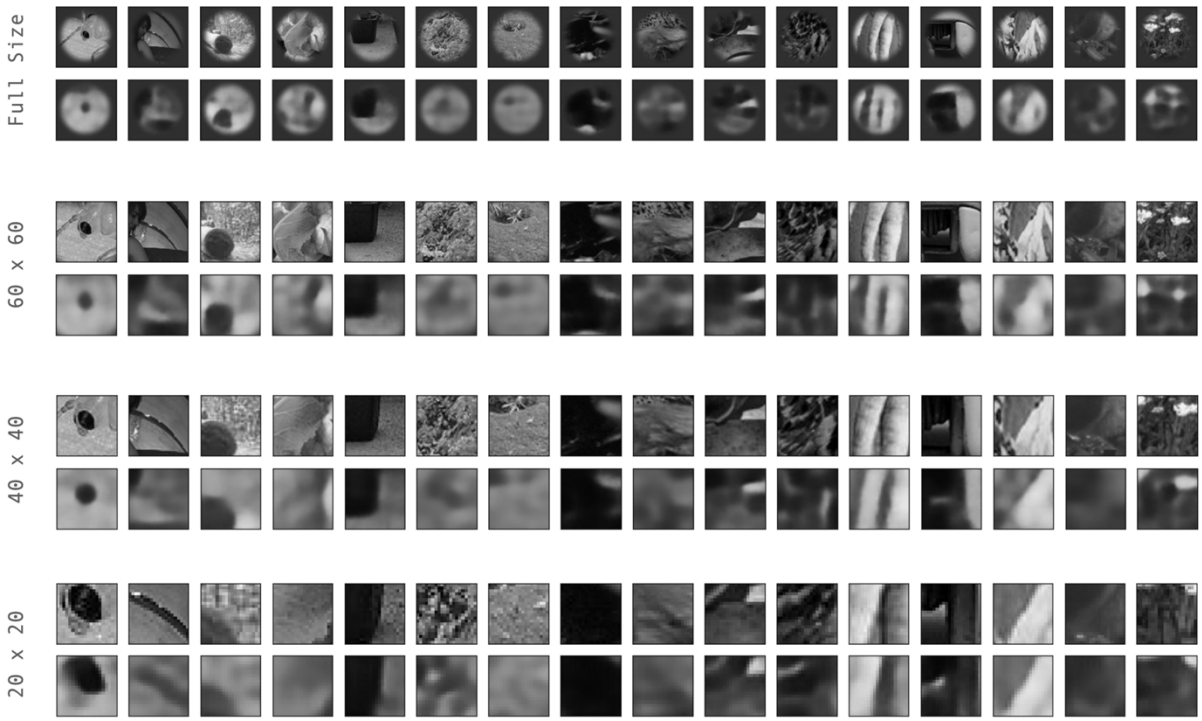

Figure 7. VAE reconstructions (2nd rows) of various size images (1st rows)

4.b) Reconstructing full images from brain signals

The second part of the decoding was mapping brain response to this latent space, to then reconstruct the images. The qualitative results of this reconstruction for the full-size images appears in Figure 8. A mean correlation of 0.5750 was observed between the reconstructed images and the original one, although this result should be considered a wide over-estimation due to the presence of the aperture contour. This contour can be trivially reconstituted by a VAE (as all images follow the same format) but is still taken into account in correlation measures. Qualitatively, the reconstructed images are usually monochromatic and very blurry. Some of them manage to reconstruct the image's overall light levels, few of them even reconstructing shape approximations, but most of the reconstructions seem to be variations of blurry squares.

4.c) Comparing Reconstruction Performance on Image Centers

For the second experiment, we observed the reconstruction performance of the network while reconstructing different center sizes of the image. The qualitative results of this reconstruction appear in Figure 8. The mean pixel-to-pixel correlations of the image centers reconstructed *from brain signals* appear on the center column below. On the right, pixel-by-pixel correlations between original images and their reconstruction using 3.b's Image VAE represent the maximum correlation *achievable* for each image size. If the ensemble network perfectly mapped brain signals to encodings, those would have been the obtained results.

Image Size	Pixel-to-Pixel correlation	Pixel-to-Pixel correlation of the corresponding Image VAE
20 x 20	0.2141	0.8145
40 x 40	0.1783	0.7695
60 x 60	0.2058	0.7599

Table 3. Pixel-to-pixel correlations of the brain-signal reconstructions and Image VAE (3b) reconstructions

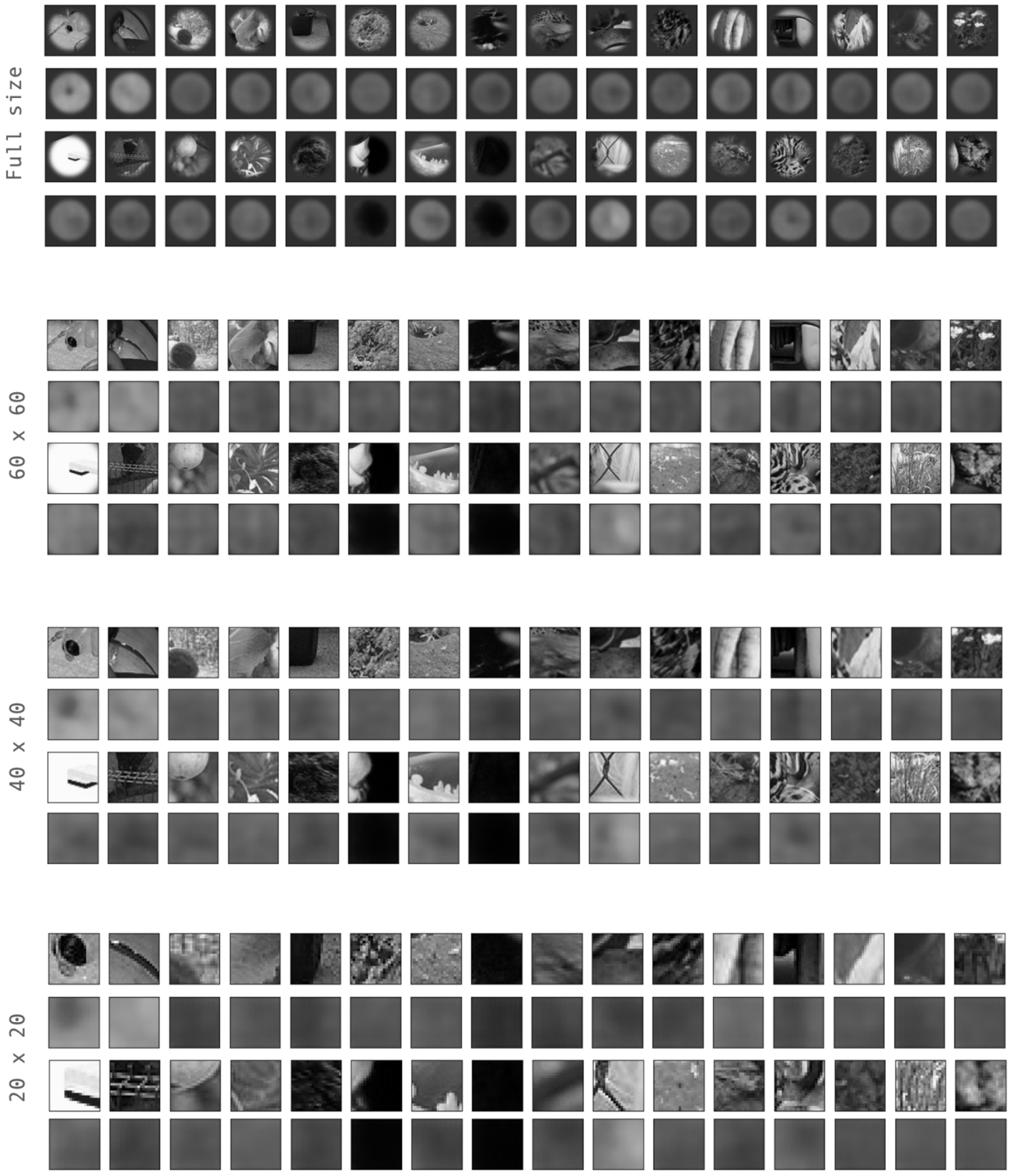

Figure 8. Reconstructions (2nd rows) of the 32 first images of the Validation Dataset (1st rows) using the mean brain responses of monkeys

5. Discussion

As seen above, the reconstruction performance of the model using only brain signals is low. The reconstructed images are not good enough for classification, yet alone for accurate reconstruction. Overall, it seems that the model generates a form of "base" image, in our case a gray square, and sometimes manages to estimate some actual image information (very basic shapes or overall color levels) that it transposes into this base generated image. In this respect, this result is very similar to what previous models have displayed when trying to reconstruct natural scenes from 2-photon calcium imaging (Zhang et al., 2022), although the nature of this "base" image can vary: for example, Zhang et al obtained a blurry mix of black and white spots, on which some features (e.g. a zebra's stripes) could be reconstructed. However, despite the low decoding results presented in this report, some remarks can be made:

- Very dark images are clearly identified as such, often being reconstructed as a black square
- In some images networks of all sizes manage to reconstruct some of the main features: the insect's eye in the first image for example.
- Overall pixel-to-pixel correlation levels at all sizes are around 0.2, showing that the model's predictions are significantly better than chance and that meaningful information is decoded from the brain signals.

On the other hand, the goal of our second experiment was to test the hypothesis that, despite only having access to the center of the image, the model would be able to predict its surround based on the same brain signals. This was proven to be true. The pixel-by-pixel correlation between original and reconstructed images was 0.2141 for 20 x 20 images, against 0.2058 for 60 x 60 images. As expected, there was no major drop in performance between

predictions covering the actual neuron's receptive field and prediction of larger image centers. These highly similar performances are coherent with our hypothesis that V1 neurons get information from their receptive field's surroundings through other connections. In some cases, the network even clearly manages to predict features that are outside of the 20 x 20 receptive field (such as the center diagonal line in image 2).

It should be noted that such results are not an irrefutable proof that neurons see the surroundings of their original receptive field: it is also possible that the model manages to infer some of the properties of the image's periphery simply based on knowing what the center looks like, rather than through information contained in brain signals. In order to prove this surrounding perception, future work could explore datasets where the center and the periphery of images have fundamentally different properties such that the latter cannot be estimated from knowing the former. If a model still achieves similar periphery performance relying on brain signals only, this would be a full confirmation that V1 neurons encode information outside of their theoretical receptive field.

In conclusion, while decoding images of natural scenes using 2-photon calcium imaging is still a difficult task, at least partially because of the procedure's noisiness, it is possible to infer some basic image information from this type of data. Furthermore, it appears that this type of single-cell recordings may allow, in the future, for the decoding of scenes far broader than these specific cells' visual fields.

6. References

Benchetrit, Y., Banville, H., & King, J.-R. (2023). *Brain decoding: toward real-time reconstruction of visual perception*. doi:10.48550/ARXIV.2310.19812

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *Journal of Neurophysiology*, *88*(5), 2530–2546. doi:10.1152/jn.00692.2001

Chen, S., Weidner, R., Zeng, H., Fink, G. R., Müller, H. J., & Conci, M. (2021). Feedback from lateral occipital cortex to V1/V2 triggers object completion: Evidence from functional magnetic resonance imaging and dynamic causal modeling. *Human Brain Mapping*, *42*(17), 5581–5594. doi:10.1002/hbm.25637

Coggan, D. D., Allen, L. A., Farrar, O. R. H., Gouws, A. D., Morland, A. B., Baker, D. H., & Andrews, T. J. (2017). Differences in selectivity to natural images in early visual areas (V1-V3). *Scientific Reports*, *7*(1), 2444. doi:10.1038/s41598-017-02569-4

Cui, Y., Qiao, K., Zhang, C., Wang, L., Yan, B., & Tong, L. (2021). GaborNet visual encoding: A lightweight region-based visual encoding model with good expressiveness and biological interpretability. *Frontiers in Neuroscience*, *15*, 614182. doi:10.3389/fnins.2021.614182

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America. A, Optics and Image Science*, *2*(7), 1160–1169. doi:10.1364/josaa.2.001160

DeAngelis, G. C., Freeman, R. D., & Ohzawa, I. (1994). Length and width tuning of neurons in the cat's primary visual cortex. *Journal of Neurophysiology*, *71*(1), 347–374. doi:10.1152/jn.1994.71.1.347

Du, B., Cheng, X., Duan, Y., & Ning, H. (2022). FMRI brain decoding and its applications in brain-computer interface: A survey. *Brain Sciences*, *12*(2), 228. doi:10.3390/brainsci12020228

Du, C., Du, C., Huang, L., & He, H. (2019). Reconstructing perceived images from human brain activities with Bayesian deep multiview learning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(8), 2310–2323. doi:10.1109/TNNLS.2018.2882456

Frégnac, Y. (2023). Flagship afterthoughts: Could the human brain project (HBP) have done better? *eNeuro*, *10*(11), ENEURO.0428-23.2023. doi:10.1523/ENEURO.0428-23.2023

Garasto, S., Bharath, A. A., & Schultz, S. R. (2018). Visual reconstruction from 2-photon calcium imaging suggests linear readout properties of neurons in mouse primary visual cortex. doi:10.1101/300392

Grienberger, C., Giovannucci, A., Zeiger, W., & Portera-Cailliau, C. (2022). Two-photon calcium imaging of neuronal activity. *Nature Reviews. Methods Primers*, *2*(1). doi:10.1038/s43586-022-00147-1

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1714), 20160102. doi:10.1098/rstb.2016.0102

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, *293*(5539), 2425–2430. doi:10.1126/science.1063736

Hayashi, R., & Kawata, H. (2018, October). Image reconstruction from neural activity recorded from monkey inferior temporal cortex using generative adversarial networks. *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Presented at the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan. doi:10.1109/smc.2018.00028

Helmchen F. Two-Photon Functional Imaging of Neuronal Activity. In: Frostig RD, editor. In Vivo Optical Imaging of Brain Function. 2nd edition. Boca Raton (FL): CRC Press/Taylor & Francis; 2009. Chapter 2. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20230/

Huang, W., Yan, H., Wang, C., Li, J., Zuo, Z., Zhang, J., … Chen, H. (2020). Perception-to-image: Reconstructing natural images from the brain activity of visual perception. *Annals of Biomedical Engineering*, *48*(9), 2323–2332. doi:10.1007/s10439-020-02502-3

Huang, W., Yan, H., Wang, C., Yang, X., Li, J., Zuo, Z., … Chen, H. (2021). Deep natural image reconstruction from human brain activity based on conditional progressively growing generative adversarial networks. *Neuroscience Bulletin*, *37*(3), 369–379. doi:10.1007/s12264-020-00613-4

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. doi:10.1038/nn1444

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. doi:10.1038/nature06713

Lee, T. S. (1996). Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(10), 959–971. doi:10.1109/34.541406

Malik, W. Q., Schummers, J., Sur, M., & Brown, E. N. (2011). Denoising two-photon calcium imaging data. *PloS One*, *6*(6), e20490. doi:10.1371/journal.pone.0020490

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-A., Morito, Y., Tanabe, H. C., … Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*(5), 915–929. doi:10.1016/j.neuron.2008.11.004

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. doi:10.1016/j.neuroimage.2010.07.073

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology: CB*, *21*(19), 1641–1646. doi:10.1016/j.cub.2011.08.031

Nurminen, L., & Angelucci, A. (2014). Multiple components of surround modulation in primary visual cortex: multiple neural circuits with multiple functions? *Vision Research*, *104*, 47–56. doi:10.1016/j.visres.2014.08.018

Saeidi, M., Karwowski, W., Farahani, F. V., Fiok, K., Taiar, R., Hancock, P. A., & Al-Juaid, A. (2021). Neural decoding of EEG signals with machine learning: A systematic review. *Brain Sciences*, *11*(11), 1525. doi:10.3390/brainsci11111525

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. A. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, *181*, 775–785. doi:10.1016/j.neuroimage.2018.07.043

Sengpiel, F., Sen, A., & Blakemore, C. (1997). Characteristics of surround inhibition in cat area 17. *Experimental Brain Research*, *116*(2), 216–228. doi:10.1007/pl00005751

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), e1006633. doi:10.1371/journal.pcbi.1006633

Shushruth, S., Ichida, J. M., Levitt, J. B., & Angelucci, A. (2009). Comparison of spatial summation properties of neurons in macaque V1 and V2. *Journal of Neurophysiology*, *102*(4), 2069–2083. doi:10.1152/jn.00512.2009

Shushruth, S., Mangapathy, P., Ichida, J. M., Bressloff, P. C., Schwabe, L., & Angelucci, A. (2012). Strong recurrent networks compute the orientation tuning of surround modulation in the primate primary visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(1), 308–321. doi:10.1523/JNEUROSCI.3789-11.2012

Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(5), 1565–1572. doi:10.1523/JNEUROSCI.4657-08.2009

Stosiek, C., Garaschuk, O., Holthoff, K., & Konnerth, A. (2003). In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(12), 7319–7324. doi:10.1073/pnas.1232232100

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, *33*(4), 1104–1116. doi:10.1016/j.neuroimage.2006.06.062

Walker, G. A., Ohzawa, I., & Freeman, R. D. (2000). Suppression outside the classical cortical receptive field. *Visual Neuroscience*, *17*(3), 369–379. doi:10.1017/s0952523800173055

Wang, P., & Cottrell, G. W. (2017). Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of Vision*, *17*(4), 9. doi:10.1167/17.4.9

Zhang, Y., Bu, T., Zhang, J., Tang, S., Yu, Z., Liu, J. K., & Huang, T. (2022). Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex. *Neural Computation*, *34*(6), 1369–1397. doi:10.1162/neco_a_01498

Zheng, X., Chen, W., Li, M., Zhang, T., You, Y., & Jiang, Y. (2020). Decoding human brain activity with deep learning. *Biomedical Signal Processing and Control*, *56*(101730), 101730. doi:10.1016/j.bspc.2019.101730