Y-Net: Fusing fMRI and Natural Scenes to Generate Neuro-Interpretable Heat Maps

Subhrasankar Chatterjee
subhrasankarphd@iitkgp.ac.in
Shreyansh Vansh Verma

Debasis Samanta

Monalisa Sarma

Indian Institute of Technology,
Kharagpur
Kharagpur, India

Abstract

Interpretability is crucial in computational neuroscience, where understanding the relationship between input stimuli and neural responses is essential. Traditional segmentation models like UNet are often seen as "black boxes," providing accurate results but lacking transparency in how they reach those results. This lack of interpretability poses a significant challenge when linking visual stimuli with the corresponding brain activity, limiting our ability to draw meaningful conclusions about the neural mechanisms underlying perception. The core issue is that conventional models do not incorporate neural data, such as fMRI, which is vital for providing contextual information about how the brain processes visual inputs. By integrating this data, the models can offer insights into the cognitive processes involved, resulting in more accurate segmentation and interpretability. To overcome this limitation, we propose the YNet architecture, which integrates fMRI-derived features with natural scene images. This fusion enhances segmentation accuracy and improves interpretability by revealing how neural activations contribute to the segmentation process. YNet allows researchers better to understand the link between visual stimuli and brain activity, offering a more transparent and insightful approach to decoding neural mechanisms of perception.

1 Introduction

In recent years, the intersection of neuroscience and computer vision has witnessed profound advancements, particularly in elucidating how visual stimuli are processed within the human brain. A compelling area of inquiry involves integrating natural scene images with functional magnetic resonance imaging (fMRI) data to uncover the neural mechanisms that underpin visual perception [3]. By combining these modalities, researchers endeavor to discern how specific features inherent in natural scenes evoke distinct patterns of brain activity, thereby enriching our comprehension of cognitive processes related to visual processing. fMRI, sensitive to changes in blood oxygenation and flow associated with neural activity, offers

insights into the brain's functional architecture by revealing activated regions in response to varied stimuli [36]. This integration of natural scene images and fMRI data presents a potent approach for investigating the intricate relationship between visual inputs and brain activity, potentially advancing neuroscientific understanding and applications in brain-computer interfaces, medical imaging, and artificial intelligence [40].

The study of visual dynamics in the brain has been a foundational pursuit in neuroscience since the seminal work of Hubel and Wiesel in the 1960s, which introduced the concept of feature detectors in the visual cortex through experiments on cats [16, 17, 18, 30]. Since then, researchers have sought to unravel the capabilities of the biological visual system. Recent decades have seen the emergence of computational models, known as encoding models, designed to predict neural responses based on stimulus inputs [22, 23, 24, 34]. Early models, such as the Gabor Wavelet Pyramid [23, 33, 34, 35] and Semantic Categorical Labelling [33, 35], relied on handcrafted feature sets tailored to specific regions of the visual cortex. However, these models needed more generalizability across different visual areas; while effective in the early stages of visual processing, they faltered in higher-order regions. The advent of deep neural networks (DNNs), particularly convolutional neural networks (CNNs), addressed this challenge by leveraging hierarchical representations [45] that align with the hierarchical organization of the visual system [14, 15, 26, 27, 51, 52]. CNNs demonstrated efficacy in predicting neural responses [8, 13, 14, 19, 22, 26, 37, 41], with early layers adept at modeling early visual areas and deeper layers proficient in capturing responses from later visual regions [2, 9, 14, 15, 25, 31, 47, 48, 49, 50, 52]. Despite their success, DNNs are often criticized for their opacity, complicating efforts to interpret their internal workings.

Interpretability in the context of neuroscience diverges significantly from that in deep learning [21]. While explainable artificial intelligence (XAI) algorithms like Guided Backpropagation and GradCAMs [38, 42, 43, 54] aim to elucidate model predictions [28, 32, 39, 44, 46], neuroscience interpretability focuses on comprehending neuronal behavior in response to stimuli. Recent efforts in neuro-interpretability [10, 11, 12, 20] have explored gradient-based approaches such as voxelwise stimulus optimization (VSO) [6], which generates region-of-interest (ROI) heatmaps based on visual study areas. However, challenges persist due to the low signal-to-noise ratio (SNR) inherent in fMRI data, which can lead to inconclusive heat maps [6]. The Dreamcatcher algorithm employs language-based encoding to generate textual interpretations from fMRI data but needs to capture the nuanced stimulus-response relationships within the biological visual system [4]. Graph-based approaches have also been proposed to map inter-region connectivity [5, 7, 8] but often neglect the stimulus' direct influence on these relationships, highlighting the need for novel methodologies that directly correlate stimulus features with neural activations.

This paper proposes the Y-Net model, designed to integrate natural scene images with corresponding fMRI data to generate heatmaps reflecting neural activity in specific visual components. The conceptual diagram of the Y-Net model is depicted in Figure 1. The training and evaluation of the Y-Net model are conducted using the Natural Scenes Dataset, a comprehensive collection of natural scene images widely used in computer vision and neuroscience research. The Y-Net model is trained with pairs of stimulus images and their corresponding fMRI scans as input, aiming to reconstruct the original stimulus image as output. This unsupervised learning approach leverages the paired nature of the dataset to learn the mapping between visual stimuli and neural responses. The performance of the proposed model is benchmarked against several standard encoding models, such as GWP, AlexNet, ResNet-50, and VOneNet, to assess the efficacy of the Y-Net model. Furthermore, layer-wise neural prediction performance is analyzed to investigate the Y-Net model's ability to

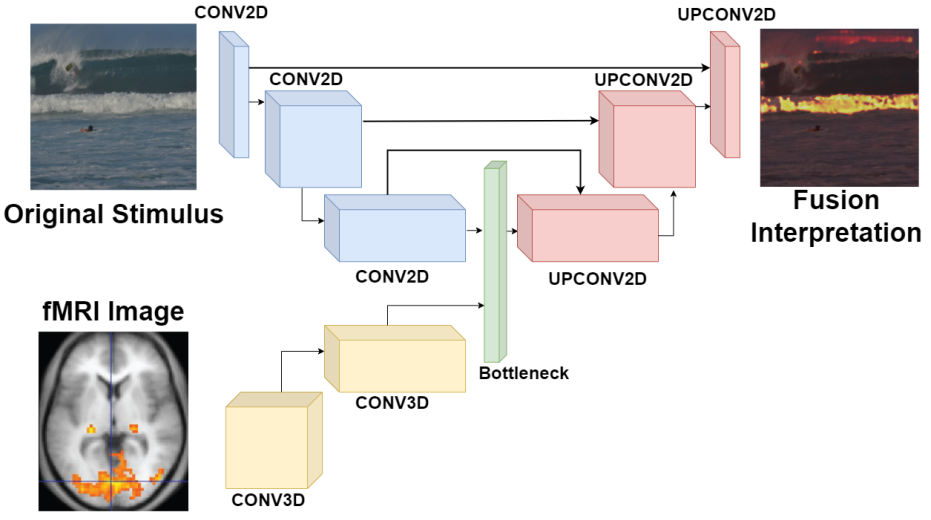

Figure 1: General Architecture of Proposed Y-Net Model. The original stimulus (image) is passed as the base of the segmentation map, where as the fMRI is passed as the neuro-informed segmatation estimator.

generalize features across its architecture. This analysis evaluates how effectively each component—encoder, latent space representation, and decoder—contributes to neural prediction accuracy. The results demonstrate that the Y-Net model achieves prediction performance comparable to the highest-performing benchmark models. Moreover, the layer-wise analysis reveals nuanced insights into feature generalization across the Y-Net architecture. Visualizations and heatmaps generated by the Y-Net model further underscore its interpretability and explanatory power.

2 Proposed Methodology

The YNet architecture is an advanced neural network designed to fuse natural scene images with fMRI data to decode and interpret the neural mechanisms underlying visual perception. This model extends the traditional UNet architecture by incorporating additional features derived from fMRI data, which enrich the representation and enhance segmentation performance. Below is a detailed description of the YNet architecture, along with its mathematical foundations.

The YNet model is composed of three primary components: an encoder, a decoder, and a feature integration module that incorporates fMRI data. The encoder is similar to the one in autoencoder, comprising several convolutional layers followed by max-pooling, along with skip residual connections to the reconstruction layers. This module extracts hierarchical features from the input image. Let x be the input image, and the encoder outputs a feature map $f(x)$ after a series of convolutional operations.

$$f(x) = \text{Encoder}(x) \tag{1}$$

The key difference in YNet is the integration of fMRI data. The fMRI data z, representing brain activation patterns corresponding to the visual stimulus, is processed through a separate

network (typically a series of fully connected layers or a convolutional neural network) to produce a feature map $f(z)$.

$$f(z) = \text{fMRI\_Network}(z) \tag{2}$$

The fMRI feature map $f(z)$ is then fused with the image feature map $f(x)$ using element-wise addition, concatenation, or a more sophisticated fusion method. If we denote the fusion operation by $\oplus$, the fused feature map F is given by:

$$F = f(x) \oplus f(z) \tag{3}$$

The fused feature map F is then passed through the decoder, which mirrors the encoder but with upsampling layers instead of pooling layers. The decoder generates the final segmentation map $\hat{y}$.

$$\hat{y} = \text{Decoder}(F) \tag{4}$$

The YNet architecture can be formalized using the following mathematical components. The basic building blocks of YNet are convolutional layers, which apply a set of filters to the input to produce feature maps. For an input I and a filter w, the convolution operation is defined as:

$$(I * w)(i,j) = \sum_m \sum_n I(i-m, j-n) w(m,n) \tag{5}$$

This operation is performed repeatedly at each layer, producing increasingly abstract representations of the input image. After each convolution, an activation function (typically ReLU) is applied to introduce non-linearity:

$$f(I) = \text{ReLU}(I * w + b) \tag{6}$$

where b is the bias term. The encoder utilizes pooling operations (typically max-pooling) to downsample the feature maps, reducing their spatial dimensions and emphasizing the most important features.

$$P(I)(i,j) = \max_{m,n} I(2i+m, 2j+n) \tag{7}$$

In the decoder, upsampling layers (e.g., transposed convolution) are used to increase the spatial dimensions of the feature maps, gradually reconstructing the spatial information.

$$U(I)(i,j) = \sum_m \sum_n I(i/m, j/n) * w(m,n) \tag{8}$$

The training of YNet is guided by a loss function that measures the discrepancy between the predicted segmentation $\hat{y}$ and the ground truth segmentation y. A common choice is the cross-entropy loss for pixel-wise classification:

$$\mathcal{L}(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i) \tag{9}$$

Alternatively, for multi-class segmentation tasks, a generalized Dice loss or a combination of Dice and cross-entropy loss can be used.

$$\mathcal{L}_{\text{Dice}}(y, \hat{y}) = 1 - \frac{2\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i} \tag{10}$$

3 Training Details

The Y-Net model is trained in an unsupervised manner, leveraging the inherent structure and relationships within the data rather than relying on explicit labels. The training process follows these detailed steps:

3.1 Dataset Description

The **Natural Scenes Dataset (NSD)** [1] includes fMRI recordings from 8 participants, who viewed 9,000–10,000 unique color natural scenes across 30–40 scan sessions, totaling 22,000–30,000 trials. Data were collected using a 7-tesla MRI scanner with a spatial resolution of 1.8 mm and a repetition time (TR) of 1.6 seconds. Stimuli, sourced from the **Microsoft COCO database** [29] and presented at 8.4° x 8.4°, were displayed for 3 seconds with 1-second intervals. A common set of 1,000 images was shown to all participants, while the rest were unique to each. Images were shown while participants fixated centrally and performed a recognition task. The data underwent temporal and spatial interpolation for preprocessing, and single-trial beta weights were computed using a general linear model (GLM) to analyze neural responses. Surface reconstructions were generated for detailed cortical characterization using Brain surface plots from Nilearn.

3.2 Data Preparation

The training dataset comprises pairs of natural scene images and corresponding fMRI scans. The fMRI data is subjected to preprocessing, including motion correction, normalization, and spatial smoothing, to enhance signal quality and minimize noise. Both the natural scene images and fMRI scans are then normalized to a common scale to facilitate effective integration. The training-testing split was considered to be 80:20. Data for 8 subjects were considered for the training and validation of the model. The generated results are calculated on the testing split averaged over 8 subjects.

3.3 Optimization

Model parameters are optimized using the Adam optimizer with an initial learning rate of 0.001. Gradient computation is performed through backpropagation, and parameter updates are carried out iteratively to minimize the composite loss function. Adam's adaptive learning rate helps in managing convergence, especially in the unsupervised setting where the model learns directly from the data without manual annotations.

4 Experiments and Results

4.1 Comparison of Y-Net Model with other standard Encoding Models

The table 1 compares the average Pearson's Correlation values, which quantify the correlation between expected and actual neural responses, for different models in the brain's left and right hemispheres. With a Pearson's $r - value$ of 0.35, GWP has a weak connection in the left hemisphere. With a value of 0.43, AlexNet performs better than this; ResNet50 advances it with 0.51. With a Pearson's $r - score$ of 0.58, VOneNet distinguishes itself with

	Mean Pearson's (r)	
Model	Left Hemisphere	Right Hemisphere
GWP	0.35 ± 0.04	0.31 ± 0.06
AlexNet	0.43 ± 0.03	0.41 ± 0.03
ResNet50	0.51 ± 0.02	0.48 ± 0.03
VOneNet	0.58 ± 0.01	0.55 ± 0.01
Proposed Y-Net	0.57 ± 0.01	0.52 ± 0.01

Table 1: This table shows a comparison of standard encoding models, namely GWP, AlexNet, ResNet50 and VOneNet agaisnt the proposed Y-net model.

significant predictive power. With a value of 0.57, the proposed Y-Net model shows almost identical accuracy and is closely behind VOneNet. The variation between VOneNet and Y-Net is not statistically significant, according to a one-tail $t-test$. With a Pearson's $r-score$ of 0.31, GWP performs poorly in the right hemisphere. With 0.41, AlexNet performs better; ResNet50 keeps improving with 0.48. Once again, leading with the highest rating of 0.55 is VOneNet; Y-Net follows closely at 0.52, barely behind VOneNet.

As confirmed by a t-test $(p < 0.01)$, the suggested Y-Net model routinely beats GWP in both hemispheres, with notable variations of 0.22 in the left and 0.21 in the right. With increases of 0.14 in the left hemisphere and 0.11 in the right, Y-Net performs noticeably better than AlexNet. With variations of 0.06 in the left hemisphere and 0.04 in the right, Y-Net likewise beats ResNet50, showcasing its excellent performance. With just minor variations of 0.01 in the left hemisphere and 0.03 in the right, Y-Net comes close behind even if VOneNet boasts the best Pearson's $r-values$. This little difference shows that Y-Net is matched in prediction accuracy and is quite competitive.

4.2 Efficacy of the representation space of Y-Net for different visual ROIs

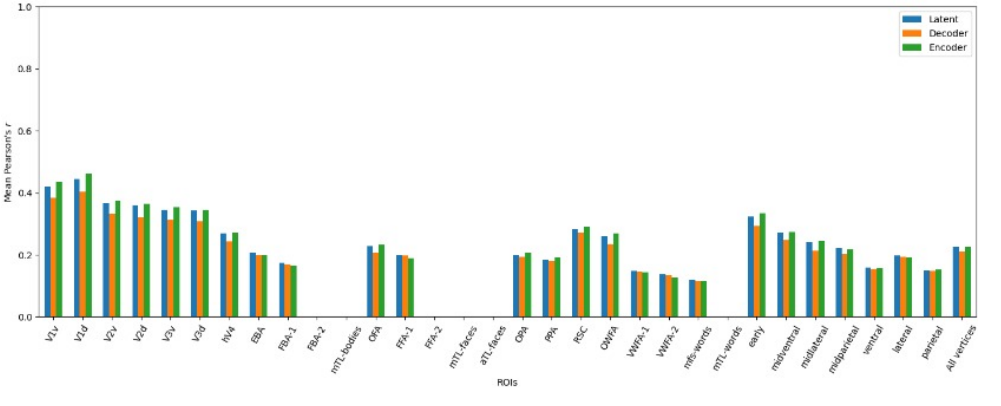

Figure 2: Comparison of extracted features from Y-Net against different visual areas. V1, V2, V3 represent the early visual areas, where as the rest represent the late visual areas.

Figure 2 compares the performance of different components (Latent, Decoder, and Encoder) across various regions of interest (ROIs) using mean Pearson's R values. These values

indicate the correlation between predicted and actual fMRI responses, measuring how well each component captures the underlying brain activity.

The encoder in the proposed model acts as a practical feature extractor, as evidenced by its consistently high correlation values across multiple ROIs. This inference suggests that the encoder is adept at capturing essential information from the input data, which is crucial for accurately predicting brain activity. The strong performance of the encoder highlights its ability to distill relevant features from the data, making it a valuable component of the overall model.

4.3 Comparison of proposed Y-Net model with standard U-Net model in terms of segmentation

Figures 3 are outputs from the proposed YNet models and pre-trained UNet. Visually comparing these images reveals a noticeable difference in the quality and accuracy of the segmentation. The second image, produced by the UNet model, exhibits a segmentation map that could be more varied and precise. The boundaries between different regions are blurred, and the segmentation lacks clarity, suggesting that the model struggles to delineate the various components within the scene accurately.

In contrast, the first image generated by the YNet model is markedly more refined. The segmentation map is cleaner, with well-defined boundaries that closely follow the structure of the original image. The regions in this output are distinct, indicating that the YNet model has effectively captured the intricate features of the image. This significant improvement in segmentation quality can be attributed to integrating fMRI-based features within the YNet model, which enhances its ability to interpret and segment complex visual scenes.

Evaluating these models using the Intersection over Union (IoU) metric further highlights the superiority of the YNet approach. The IoU measures the overlap between the predicted segmentation and the ground truth, providing a quantitative assessment of segmentation accuracy. The UNet model, with its noisier output, achieves a lower IoU (0.34) score due to the poor alignment between the predicted and actual regions. On the other hand, the YNet model, with its more precise and accurate segmentation, achieves a higher IoU (0.77) score, reflecting its ability to match the expected segmentation closely.

4.4 Abalation Study of the Y-Net Model

In the proposed YNet model, the fMRI images are fed into a subsidiary input layer, which generates features and adds up to the latent space, as in visible in figure 4. This feature extraction mechanism allows the model to generate a much better feature vector for the latent space. By enhancing the quality of the feature vector that is passed to the latent space, the model ensures that the decoder receives more precise and informative inputs. As a result, the decoder in the proposed YNet model can outperform the standard UNet segmentation output. The improved feature vector in the latent space enables the decoder to reconstruct the fMRI data more accurately. This results demonstrates that the synergy between the encoder's feature extraction and the subsequent linear regression leads to a more robust model, ultimately enhancing the decoder's performance in generating superior outputs compared to traditional methods.

Including fMRI features in the YNet model is crucial for improving segmentation performance. fMRI data provides additional neural activation information corresponding to different visual stimuli, allowing the model better to understand the underlying structure and

Figure 3: Comparison of segmentation maps (Y-Net , on the left, vs U-Net , on the right). It can be observed that the generated results from Y-Net has a better representation in comparison to pre-trained U-Net

context of the image. This enriched information helps the YNet model make more informed decisions during segmentation, leading to more accurate and reliable outputs.

4.5 Interpretability of the generated fusion map

As observed from the previous results, the neuro-informed segmentation map captures the object in the context relevant to the fMRI data. The objects under a higher level of cognition can be identified using a mechanism called binning. Binning is a technique that reduces noise and enhances segmentation by grouping the pixel values into a specified number of bins. The majorly relevant segmented areas within the image can be identified by binning the pixel values corresponding to the fMRI data's voxel firing. The histogram displayed represents

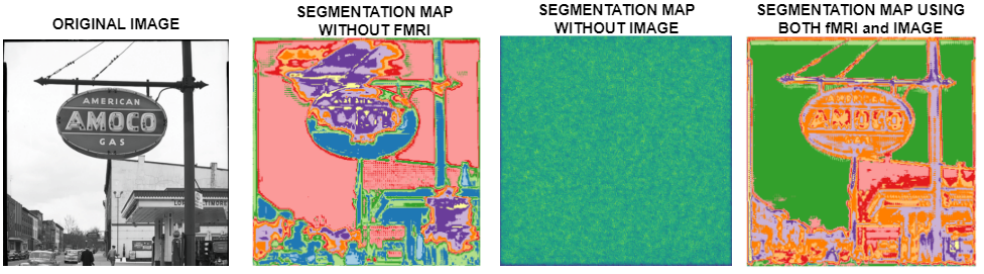

Figure 4: Abalation study of the components of Y-Net. Without the fMRI input, the Y-Net behaves exactly as the U-Net model with cluttered segmentation map. Without the base image, the produced map is just a random noise. On the contrary combining the two produces a reasonably good segmentation for input images.

the distribution of pixel values in a segmented output image. This type of analysis helps identify the pixel ranges corresponding to the image's different segments. Each segment has a specific range of pixel values, and these ranges are represented by the peaks in the histogram, achieving the concept of binning for identifying the interpretable segmentation map. A detailed result can be observed in figure 5.

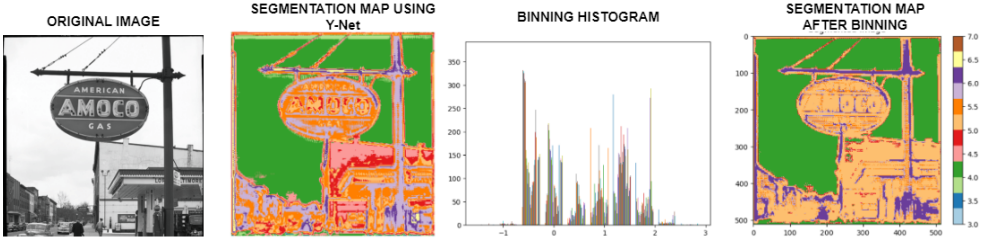

Figure 5: Interpretability map of the stimulus using Y-Net model and binning. The binned segmentation map generates a clear interpretation of the objects more relevant to the cognitive processes captured by the fMRI.

5 Conclusion

The YNet architecture effectively enhances image segmentation by integrating fMRI-derived features, allowing it to decode visual stimuli more accurately. By combining traditional convolutional layers with neural data, YNet captures both the visual and cognitive aspects of the input, leading to superior segmentation performance. The mathematical foundations of YNet provide a robust framework for its operation, ensuring precise and reliable results. This approach outperforms traditional models like UNet, making YNet a powerful tool for decoding and interpreting complex visual scenes.

References

[1] Emily Allen, Ghislain St-Yves, Yihan Wu, Jesse Breedlove, Jacob Prince, Logan Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25, 01 2022. doi: 10.1038/s41593-021-00962-x.

[2] Pouya Bashivan, Kohitij Kar, and James Dicarlo. Neural population control via deep image synthesis, 11 2018.

[3] Rosa Cao and Daniel Yamins. Explanatory models in neuroscience: Part 1 – taking mechanistic abstraction seriously, 04 2021.

[4] Subhrasankar Chatterjee and Debasis Samanta. Dreamcatcher: Revealing the language of the brain with fmri using gpt embedding. *arXiv preprint arXiv:2306.10082*, 2023.

[5] Subhrasankar Chatterjee and Debasis Samanta. Enhancing graph-based representation learning with adversarial policy gradient: A hyperparameter analysis. In Dipak Kumar Kole, Shubhajit Roy Chowdhury, Subhadip Basu, Dariusz Plewczynski, and Debotosh Bhattacharjee, editors, *Proceedings of 4th International Conference on Frontiers in Computing and Systems*, pages 307–320, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-97-2611-0.

[6] Subhrasankar Chatterjee and Debasis Samanta. A gradient-based approach to interpreting visual encoding models. In Harkeerat Kaur, Vinit Jakhetiya, Puneet Goyal, Pritee Khanna, Balasubramanian Raman, and Sanjeev Kumar, editors, *Computer Vision and Image Processing*, pages 331–342, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-58181-6.

[7] Subhrasankar Chatterjee, Subrata Pain, and Debasis Samanta. A novel graph representation learning approach for visual modeling using neural combinatorial optimization. In *Pattern Recognition and Machine Intelligence*, pages 228–237, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-45170-6.

[8] Subhrasankar Chatterjee, Subrata Pain, and Debasis Samanta. Adversarial policy gradient for learning graph-based representation in human visual processing, 2023. URL https://openreview.net/forum?id=5-ROmmBJKV.

[9] Radoslaw Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, 06 2016. doi: 10.1038/srep27755.

[10] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023. doi: 10.1101/2022.03.28.485868. URL https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868.

[11] Katharina Dobs. Using deep neural networks to test possible origins of human face perception, 08 2023.

[12] Katharina Dobs, Julio Martinez, Alexander Kell, and Nancy Kanwisher. Brain-like functional specialization emerges spontaneously in deep neural networks. *Science advances*, 8:eabl8913, 03 2022. doi: 10.1126/sciadv.abl8913.

[13] Adrien Doerig, Rowan Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace Lindsay, Konrad Kording, Talia Konkle, Marcel Gerven, Nikolaus Kriegeskorte, and Tim Kietzmann. The neuroconnectionist research programme. *Nature reviews. Neuroscience*, 24, 05 2023. doi: 10.1038/s41583-023-00705-w.

[14] Michael Eickenberg, Alexandre Gramfort, Gael Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152, 10 2016. doi: 10.1016/j.neuroimage.2016.10.001.

[15] Umut Güçlü and Marcel van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 35:10005–10014, 07 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

[16] D HUBEL and T WIESEL. Receptive fields of single neurons in the cat's striate cortex. *The Journal of physiology*, 148:574–91, 11 1959. doi: 10.1113/jphysiol.1959.sp006308.

[17] D HUBEL and T WIESEL. Receptive fields, binocular interaction and functional architectures in cats visual cortex. *The Journal of physiology*, 160:106–54, 02 1962. doi: 10.1113/jphysiol.1962.sp006837.

[18] J.P. Jones and L.A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58:1233–58, 01 1988. doi: 10.1152/jn.1987.58.6.1233.

[19] Nancy Kanwisher, Meenakshi Khosla, and Katharina Dobs. Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, 46, 01 2023. doi: 10.1016/j.tins.2022.12.008.

[20] Nancy Kanwisher, Meenakshi Khosla, and Katharina Dobs. Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, 46, 01 2023. doi: 10.1016/j.tins.2022.12.008.

[21] Kohitij Kar, Simon Kornblith, and Evelina Fedorenko. Interpretability of artificial neural network models in artificial intelligence vs. neuroscience, 06 2022.

[22] Kendrick Kay. Principles for models of neural information processing. *NeuroImage*, 180, 08 2017. doi: 10.1016/j.neuroimage.2017.08.016.

[23] Kendrick Kay, Thomas Naselaris, Ryan Prenger, and Jack Gallant. Identifying natural images from human brain activity. *Nature*, 452:352–5, 04 2008. doi: 10.1038/nature06713.

[24] Kendrick Kay, Jonathan Winawer, Aviv Mezer, and Brian Wandell. Compressive spatial summation in human visual cortex. *Journal of neurophysiology*, 110, 04 2013. doi: 10.1152/jn.00105.2013.

[25] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10:e1003915, 11 2014. doi: 10.1371/journal.pcbi.1003915.

[26] Nikolaus Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1: 417–446, 11 2015. doi: 10.1146/annurev-vision-082114-035447.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL https://doi.org/10.1145/3065386.

[28] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10:e0130140, 07 2015. doi: 10.1371/journal.pone.0130140.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Zitnick. Microsoft coco: Common objects in context. volume 8693, 04 2014. ISBN 978-3-319-10601-4. doi: 10.1007/978-3-319-10602-1_48.

[30] Margaret Livingstone. Mechanisms of direction selectivity in macaque v1. *Neuron*, 20: 509–26, 03 1998. doi: 10.1016/S0896-6273(00)80991-5.

[31] Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115:201719616, 08 2018. doi: 10.1073/pnas.1719616115.

[32] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 05 2017. doi: 10.1016/j.patcog.2016.11.008.

[33] Thomas Naselaris, Ryan Prenger, Kendrick Kay, Michael Oliver, and Jack Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63:902–15, 09 2009. doi: 10.1016/j.neuron.2009.09.006.

[34] Thomas Naselaris, Kendrick Kay, Shinji Nishimoto, and Jack Gallant. Encoding and decoding in fmri. *NeuroImage*, 56:400–10, 05 2011. doi: 10.1016/j.neuroimage.2010.07.073.

[35] Shinji Nishimoto, An Vu, Thomas Naselaris, Yuval Benjamini, B. Yu, and Jack Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology : CB*, 21:1641–6, 09 2011. doi: 10.1016/j.cub.2011.08.031.

[36] Hirotaka Onoe. Advanced neurocircuit mapping via non-invasive magnetic resonance imaging techniques. *Brain and nerve = Shinkei kenkyū no shinpo*, 76:821–826, 07 2024. doi: 10.11477/mf.1416202689.

[37] Blake Richards, Timothy Lillicrap, Philippe Beaudoin, Y. Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Costa, Archy Berker, Surya Ganguli, Colleen Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace Lindsay, Kenneth Miller, Richard Naud, Christopher Pack, and Konrad Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22:1761–1770, 11 2019. doi: 10.1038/s41593-019-0520-2.

[38] Ramprasaath Rs, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128, 02 2020. doi: 10.1007/s11263-019-01228-7.

[39] Wojciech Samek, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2660–2673, 11 2017. doi: 10.1109/TNNLS.2016.2599820.

[40] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib Majaj, Rishi Rajalingham, Elias Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel Yamins, and James Dicarlo. Brain-score: Which artificial neural network for object recognition is most brain-like?, 09 2018.

[41] Thomas Serre. Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, 5, 09 2019. doi: 10.1146/annurev-vision-091718-014951.

[42] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. 05 2016.

[43] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. 04 2017.

[44] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *preprint*, 12 2013.

[45] Fabián Soto and F. Ashby. *Encoding Models in Neuroimaging*, pages 421–472. 04 2023. ISBN 9781108830676. doi: 10.1017/9781108902724.011.

[46] Jost Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. 12 2014.

[47] Ghislain St-Yves and Thomas Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces, 04 2017.

[48] Katherine Storrs, Tim Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human it well, after training and fitting. *Journal of Cognitive Neuroscience*, 33:1–21, 07 2021. doi: 10.1162/jocn_a_01755.

[49] Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Scientific Reports*, 8, 02 2018. doi: 10.1038/s41598-018-22160-9.

[50] Will Xiao and Gabriel Kreiman. Xdream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLOS Computational Biology*, 16:e1007973, 06 2020. doi: 10.1371/journal.pcbi.1007973.

[51] Daniel Yamins and James Dicarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19:356–365, 02 2016. doi: 10.1038/nn.4244.

[52] Daniel Yamins, Ha Hong, Charles Cadieu, Ethan Solomon, Darren Seibert, and James Dicarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 05 2014. doi: 10.1073/pnas.1403112111.

[53] Yiming Zhang, Ying Hu, Xiongkuo Min, Yan Zhou, and Guangtao Zhai. fmri exploration of visual quality assessment. *arXiv preprint arXiv:2404.18162*, 2024.

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. 12 2016. doi: 10.1109/CVPR.2016.319.