Estimating pRF and fLoc from Natural Images using Integrated Gradient Correlation

Pierre Lelièvre — Chien-Chung Chen

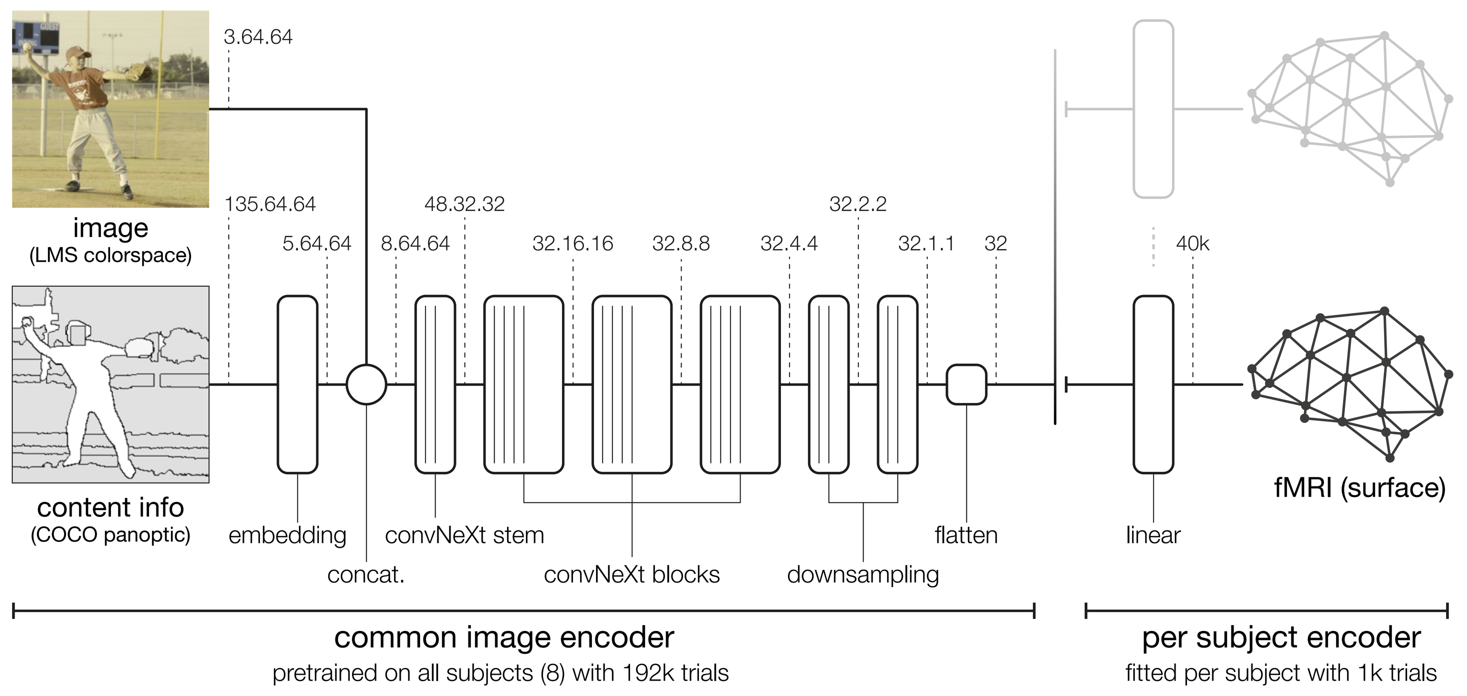
Department of Psychology, National Taiwan University, Taipei, TAIWAN

Objectives

- to develop an integrated method for pRF and fLoc estimation
- to provide information with a wider validity than artificial stimuli by the use of natural images
- to increase the encoding accuracy of brain data with deep models
- to extend the notion of pRF beyond luminance contrast stimuli

Deep Encoding Model

- Our model encodes fMRI data from visual stimuli — Natural Scenes Dataset (NSD)[1].
- It uses a partially pre-trained procedure, that requires only 1k images for new subjects.



IGC : Integrated Gradient Correlation

- IGC[•] is a dataset-wise attribution method making deep models interpretable by revealing the localization of input information relevant to output predictions at a task-level.
- It extends existing path attribution methods, e.g., Integrated Gradients[2] (IG).
- IGC attributions $\boldsymbol{b}$ are defined as:

scaled covaviance

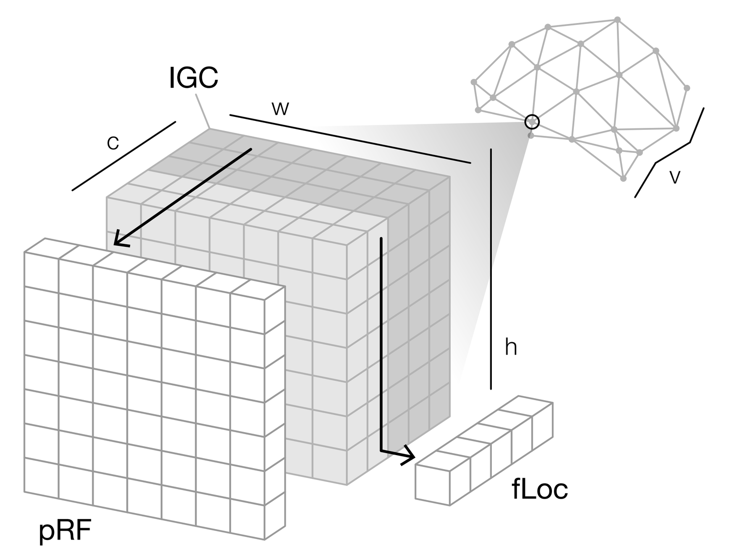
for the j-th component

$$\boldsymbol{b}_j = \frac{1}{\sigma_{f(\mathbf{x})}\sigma_{\mathrm{y}}} \mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}} \left[\boldsymbol{a}_j^{(i)} \times \left(y^{(i)} - \mu_{\mathrm{y}}\right)\right]$$

diff. of true output $\boldsymbol{y}^{(i)}$ and the mean of all $\mathbf{y}$

SD of model f predictions — SD of true outputs y — dataset $\mathcal{D}$ — IG attributions for input $\boldsymbol{x}^{(i)}$

- IGC attributions can be summarized over ROI $\mathcal{R}$ by summation: $$\boldsymbol{b}_{\mathcal{R}} = \sum_{j\in\mathcal{R}} \boldsymbol{b}_j$$
- The total attribution sums to a prediction score, i.e. the correlation between model predictions $f(\mathbf{x})$ and true outputs y: $$\sum_{j=1}^{m} \boldsymbol{b}_j = \rho\left(f(\mathbf{x}), \mathrm{y}\right)$$

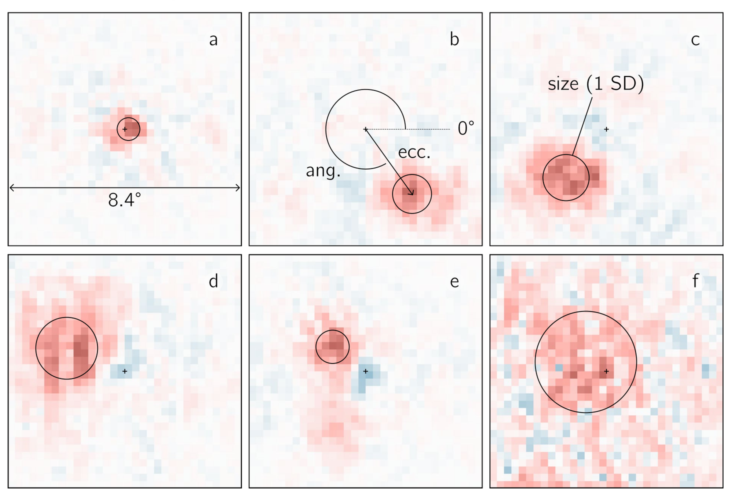


- Applied on our deep encoding model, IGC attributions have shape (c,h,w) for each vertex of the brain surface (v).
- A summation over channels c (LMS + COCO cat.) directly provides pRF maps.
- A summation over spatial dimensions h,w exposes the predictive power of each category, actually reflecting functional specifications (fLoc).

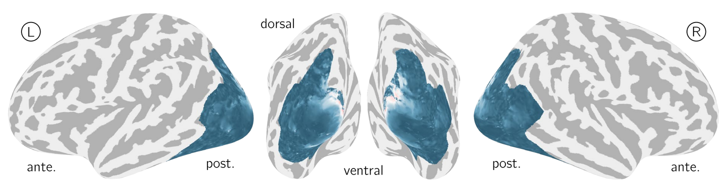
However,
- IGC is not restricted to fMRI encoding or decoding inquiries.
- IGC is applicable to any deep model architecture and data.

pRF : population Receptive Field

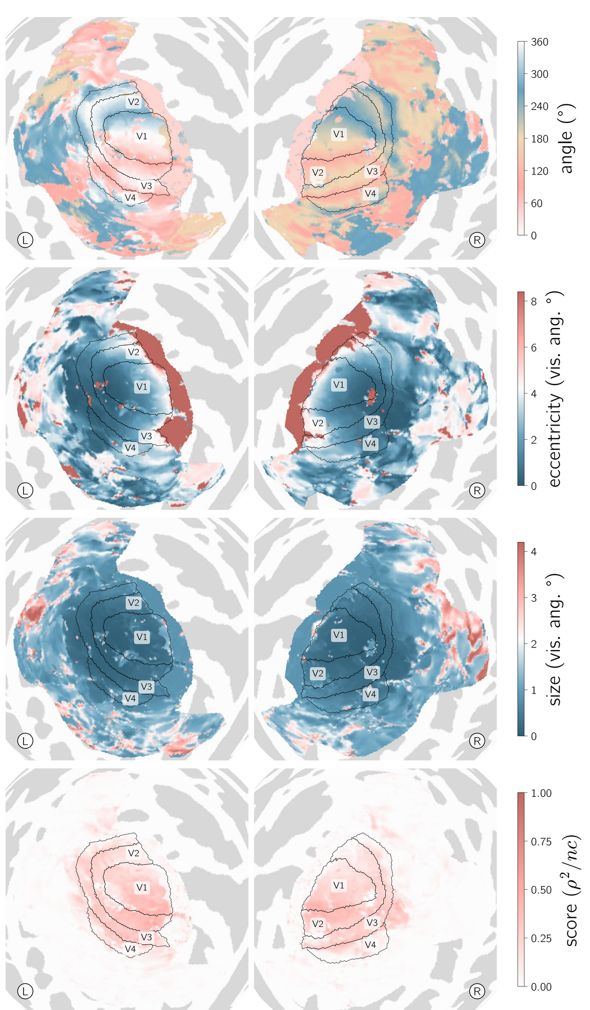
- IGC maps with fitted Gaussians



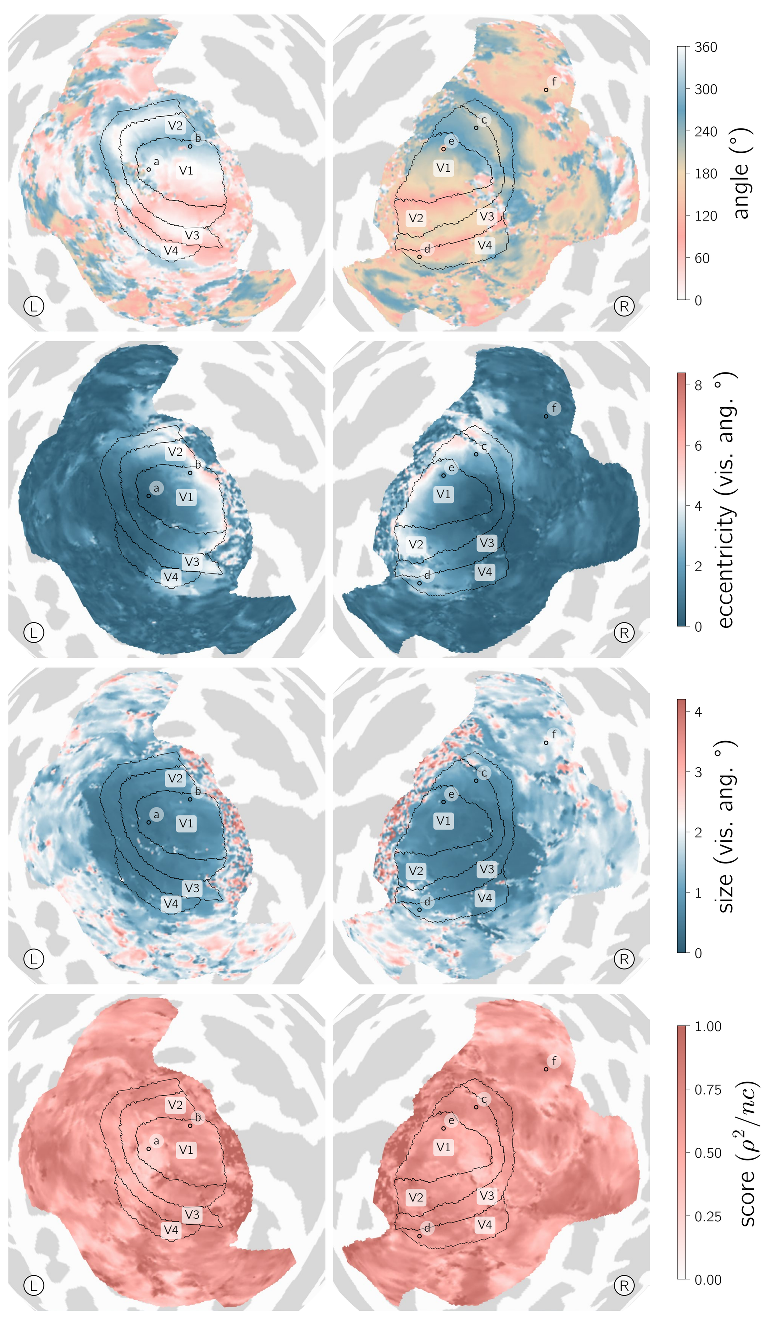
- visual cortex ROI

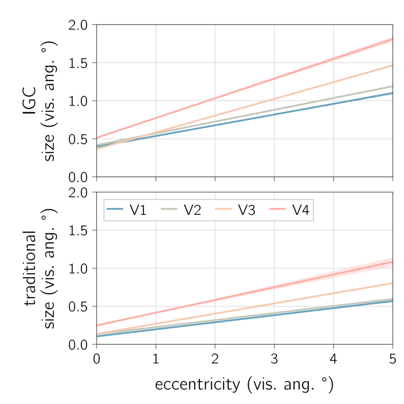


- traditional pRF
  - compressive model from Kay et al.[3]

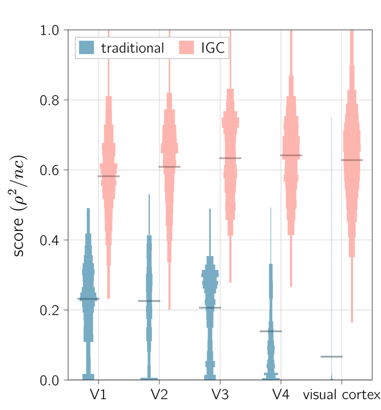


- pRF from IGC
  - simple 2-d Gaussian model





- eccentricity — size proportionality verification

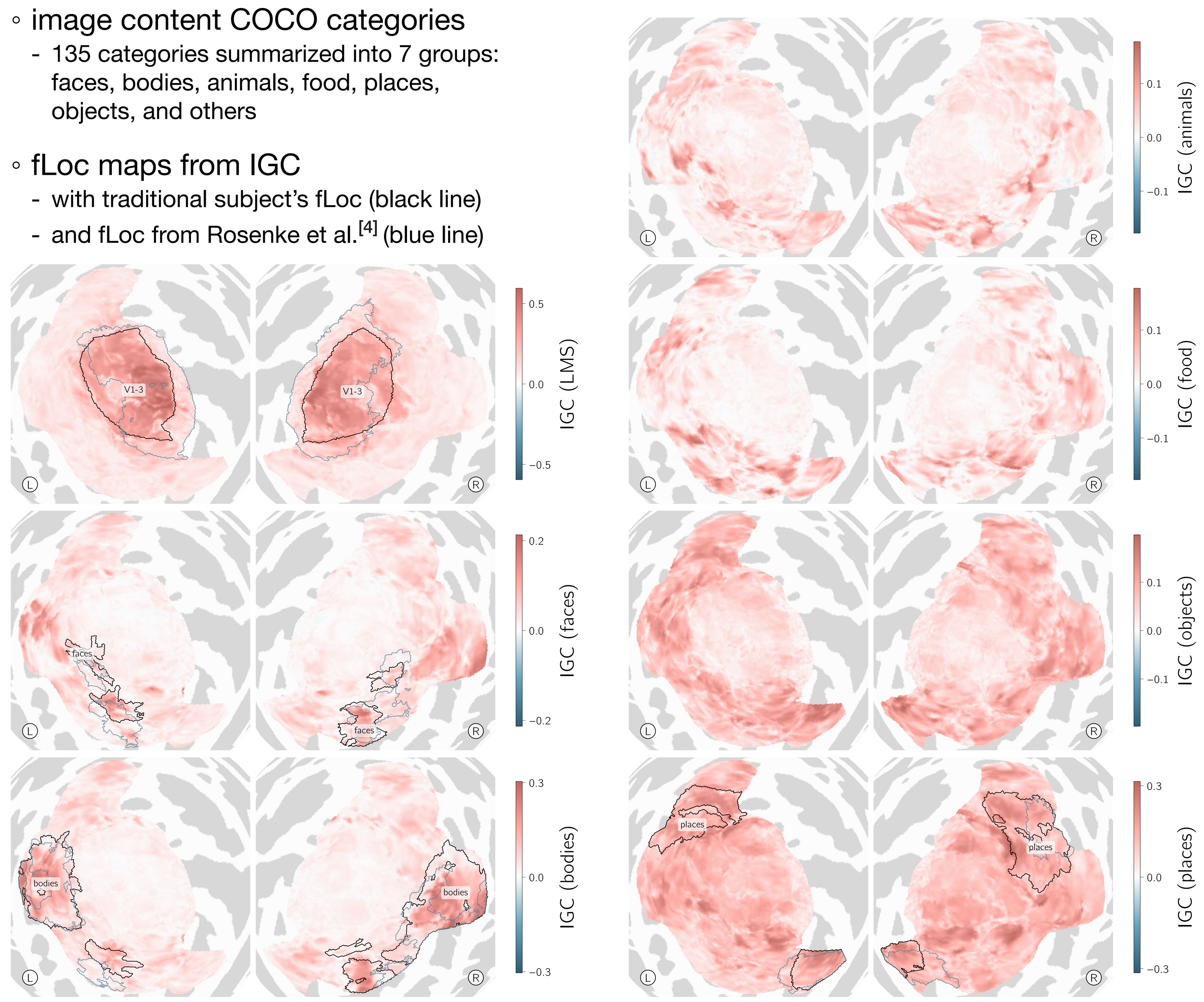


encoding accuracy

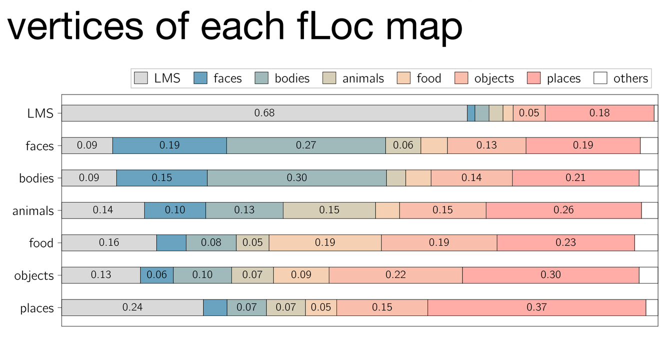
- Our deep encoding model outperforms existing pRF models by several folds on natural images in terms of variance explained.
- pRF estimation from IGC is coherent and presents fewer artifacts.

fLoc : functional Localizer

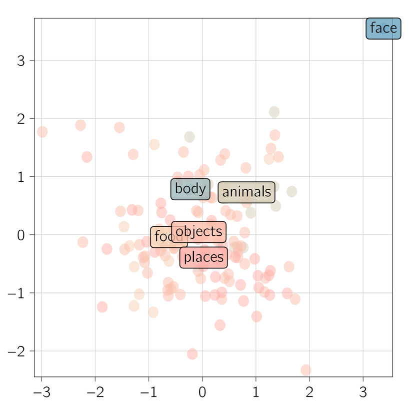
- image content COCO categories
  - 135 categories summarized into 7 groups: faces, bodies, animals, food, places, objects, and others
- fLoc maps from IGC
  - with traditional subject's fLoc (black line)
  - and fLoc from Rosenke et al.[4] (blue line)



- IGC attributions ratio of top-5% vertices of each fLoc map



- learned category embedding
  - 2-d PCA from original 5-d embeddings
  (exp. var. 0.99, 0.84)



- IGC fLoc directly reflect the predictive power of image content data.
- It avoids artificially sharp boundaries of contrastive statistical analysis.
- However, the large overlap of functional attributions prompts us to rethink higher level visual area interactions with further investigations.
- fLoc from IGC are easily extensible to more diverse or finer grain localizers of any content present in natural images.

poster — arXiv — git

Contact contact@plelievre.com

Acknowledgment

Supported by 113-2811-H-002-037-MY3 from the National Science and Technology Council (NSTC)

References

[•] Lelièvre, P., & Chen, C.-C. (2024). Integrated Gradient Correlation: A Dataset-wise Attribution Method. arXiv:2404.13910

[1] Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature Neuroscience, 25(1), 116–126.

[2] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks (No. arXiv:1703.01365). arXiv. http://arxiv.org/abs/1703.01365

[3] Kay, K. N., Winawer, J., Mezer, A., & Wandell, B. A. (2013). Compressive spatial summation in human visual cortex. Journal of Neurophysiology, 110(2), 481–494.

[4] Rosenke, M., Van Hoof, R., Van Den Hurk, J., Grill-Spector, K., & Goebel, R. (2021). A Probabilistic Functional Atlas of Human Occipito-Temporal Visual Cortex. Cerebral Cortex, 31(1), 603–619.