Inverse receptive field attention for naturalistic image reconstruction from the brain

Lynn Le (lynn.le@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

Thirza Dado (t.dado@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

Katja Seeliger (k.seeliger@donders.ru.nl)
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Paolo Papale (p.papale@nin.knaw.nl)
Netherlands Institute for Neuroscience, Amsterdam, Netherlands

Antonio Lozano (a.lozano@nin.knaw.nl)
Netherlands Institute for Neuroscience, Amsterdam, Netherlands

Pieter Roelfsema (p.roelfsema@nin.knaw.nl)
Netherlands Institute for Neuroscience; Department of Integrative Neurophysiology, Centre for Neurogenomics and Cognitive Research, Vrije Universiteit; Department of Psychiatry, Amsterdam UMC, University of Amsterdam; Laboratory of Visual Brain Therapy, Paris, France

Yağmur Güçlütürk (y.gucluturk@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

Marcel van Gerven (marcel.vangerven@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

Umut Güçlü (u.guclu@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

Abstract

Visual perception in the brain largely depends on the organization of neuronal receptive fields. Although extensive research has delineated the coding principles of receptive fields, most studies have been constrained by their foundational assumptions. Moreover, while machine learning has successfully been used to reconstruct images from brain data, this approach faces significant challenges, including inherent feature biases in the model and the complexities of brain structure and function. In this study, we introduce an inverse receptive field attention (IRFA) model, designed to reconstruct naturalistic images from neurophysiological data in an end-to-end fashion. This approach aims to elucidate the tuning properties and representational transformations within the visual cortex. The IRFA model incorporates an attention mechanism that determines the inverse receptive field for each pixel, weighting neuronal responses across the visual field and feature spaces. This method allows for an examination of the dynamics of neuronal representations across stimuli in both spatial and feature dimensions. Our results show highly accurate reconstructions of naturalistic data, independent of pre-trained models. Notably, IRF models trained on macaque V1, V4, and IT regions yield remarkably consistent spatial receptive fields across different stimuli, while the features to which neuronal representations are selective exhibit significant variation. Additionally, we propose a data-driven method to explore representational clustering within various visual areas, further providing testable hypotheses.

Keywords: neural decoding; image reconstruction; receptive field; attention; brain-computer interface

Introduction

Perception relies on selective attention to spatial and feature-based cues (Desimone & Duncan, 1995; Posner & Petersen, 1990; Serences & Yantis, 2006). Neural decoding models have leveraged this principle to reconstruct visual stimuli from brain data, but many rely on external generative models or spatial priors, limiting biological interpretability (Shen, Horikawa, Majima, & Kamitani, 2019; Takagi & Nishimoto, 2023).

We present Inverse Receptive Field Attention (IRFA), an end-to-end model that reconstructs naturalistic images from macaque V1, V4, and IT recordings. IRFA learns spatial and feature-based inverse receptive fields (IRFs) for each image pixel, without pretrained networks or retinotopic assumptions. This disentangled attention mechanism yields accurate reconstructions and reveals interpretable visual representations across cortical areas (Le et al., 2022, 2025).

Preliminaries

Our goal is to reconstruct visual stimuli S from multi-unit activity R recorded in visual cortex, via a decoding function $S = d(R)$.

While classical receptive fields (RFs) describe which image regions influence neural activity (Hubel & Wiesel, 1959), inverse receptive fields (IRFs) indicate which neurons contribute to each image pixel. IRFA estimates these IRFs using attention-based mappings over spatial (electrode) and feature (channel) dimensions. This provides a biologically interpretable link between distributed neural codes and pixel-wise reconstructions.

Methods

IRFA maps neural activity from 15 electrode arrays (V1, V4, IT) to a virtual grid with learned feature channels, capturing both spatial and feature selectivity. Inputs are processed via multi-head embeddings with positional encodings in x and y, then passed through a transpose attention module that assigns spatial and feature weights per pixel.

These attention weights guide a U-NET decoder trained to reconstruct 96×96 color images. The model is optimized end-to-end using an L1 loss, VGG-based perceptual loss, and a 5-layer adversarial discriminator, co-trained over 400 epochs with Adam.

Each pixel is assigned an interpretable inverse receptive field (IRF): a soft mapping over electrode channels (spatial) and feature maps. IRFs emerge directly from reconstruction loss, without priors, and determine how neural activity modulates image reconstruction.

Results and Discussion

More feature channels improve reconstruction. Increasing the number of learned feature channels from 4 to 64 significantly improved reconstruction quality (Fig. 2). Richer feature representations consistently outperformed simpler variants and exceeded a strong end-to-end baseline (Shen, Dwivedi, Majima, Horikawa, & Kamitani, 2019).

Biological tuning emerges from data. Without retinotopy maps or pretrained image models, IRFA uncovers known principles of cortical organization. Spatial maps mirror retinotopy, and flexible feature tuning aligns with the biased competition theory of attention (Desimone & Duncan, 1995).

Consistent spatial IRFs and dynamic feature tuning.

Spatial IRFs emerged naturally and remained consistent across stimuli for each of the 15 electrode channels (Fig. 3A,B). V1 IRFs were more localized than those in V4 and IT, reflecting cortical hierarchy. This suggests IRFA recovers biologically plausible spatial organization directly from data.

We quantified IRF variability across stimuli using standard deviation per electrode. Spatial IRFs showed low variance, while feature IRFs varied significantly (Fig. 3B), indicating stimulus-dependent feature attention. IRFA thus captures both stable spatial structure and dynamic feature selectivity.

Interpretability without generative priors. Unlike diffusion-based models that emphasize realism, IRFA offers grounded reconstructions and interpretable IRFs. It builds on Brain2Pix (Le et al., 2022) and improves accuracy through attention disentanglement.

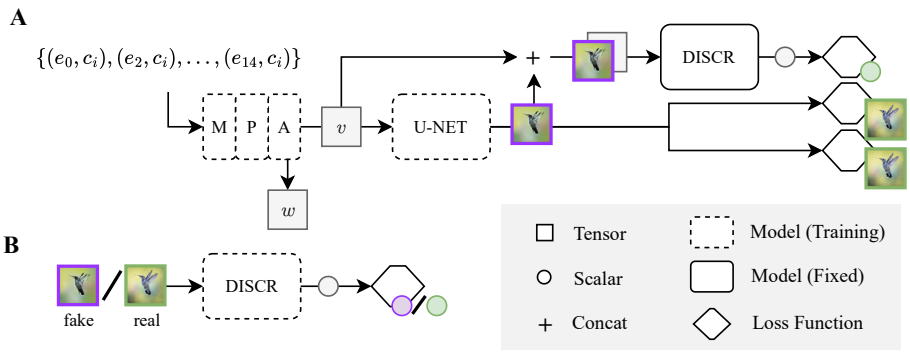

Figure 1: Schematic of our IRF reconstruction model comprising two forward passes and training five components simultaneously. **A**: Input from 15 electrode arrays—7 from V1, 4 from V4, and 4 from IT, with n channels each—is processed through multi-head embedding (M), positional encoding (P), and transpose attention (A) to create an attention map w and a weighted map v. The map v is used by the U-NET for naturalistic image reconstruction, evaluated using discriminator, VGG, and L1 losses. The map w will be analyzed for visualizing the spatial IRFs and feature IRFs. **B**: The second forward pass involves discriminator training on U-NET 'fake' outputs versus 'real' stimuli.

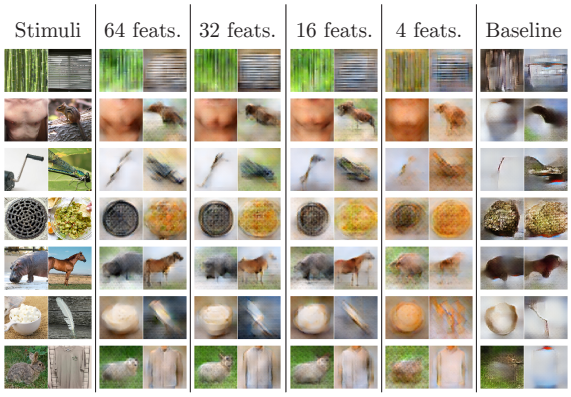

Figure 2: Reconstructions from models trained with different numbers of feature channels.

Summary

We introduce Inverse Receptive Field Attention (IRFA), an end-to-end model that reconstructs naturalistic images from macaque neural activity while learning spatial and feature-based inverse receptive fields (IRFs). IRFA assigns each pixel dynamic attention weights over neural inputs, disentangling spatial and feature selectivity without pretrained vision models or spatial priors. Our model yields accurate reconstructions and reveals biologically grounded organization: spatial IRFs emerge consistently across stimuli (Fig. 3A), reflecting known retinotopy, while feature IRFs remain flexible. These findings demonstrate that IRFA can recover meaningful neural representations directly from data, offering a testable, interpretable framework for visual decoding.

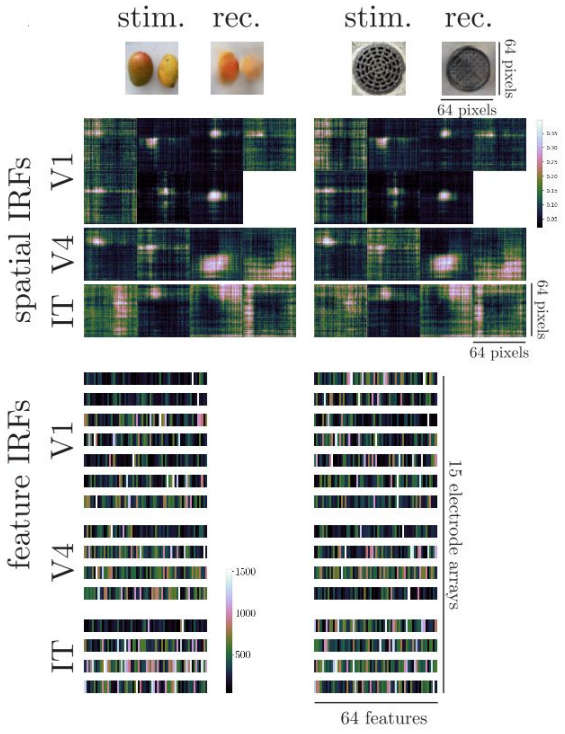

Figure 3: **A.** Learned spatial (top) and feature (bottom) IRFs assigned per reconstructed test image.

References

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*, 574-591. doi: 10.1113/jphysiol.2009.174151

Le, L., Ambrogioni, L., Seeliger, K., Güçlütürk, Y., van Gerven, M., & Güçlü, U. (2022). Brain2pix: Fully convolutional naturalistic video frame reconstruction from brain activity. *Frontiers in Neuroscience*, *16*, 940972.

Le, L., Papale, P., Seeliger, K., Lozano, A., Dado, T., Wang, F., ... Güçlü, U. (2025). Monkeysee: Space-time-resolved reconstructions of natural images from macaque multi-unit activity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems.*

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*(1), 25–42.

Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences*, *10*(1), 38–45.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., & Kamitani, Y. (2019). End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, *13*, 432276.

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

Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14453–14463).