Neural encoding with affine feature response transforms

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

Nils Kimman (n.kimman@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

Thirza Dado (thirza.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

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

Yağmur Güçlütürk (y.gucluturk@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

Current linearizing encoding models that predict neural responses to sensory input typically neglect neuroscience-inspired constraints that could enhance model efficiency and interpretability. To address this, we propose a new method called affine feature response transform (AFRT), which exploits the brain's retinotopic organization. Applying AFRT to encode multi-unit activity in areas V1, V4, and IT of the macaque brain, we demonstrate that AFRT reduces redundant computations and enhances the performance of current linearizing encoding models by segmenting each neuron's receptive field into an affine retinal transform, followed by a localized feature response. Remarkably, by factorizing receptive fields into a sequential affine component with three interpretable parameters (for shifting and scaling) and response components with a small number of feature weights per response, AFRT achieves encoding with orders of magnitude fewer parameters compared to unstructured models. We show that the retinal transform of each neuron's encoding agrees well with the brain's receptive field. Together, these findings suggest that this new subset within spatial transformer network can be instrumental in neural encoding models of naturalistic stimuli.

Keywords: Neural encoding; convolutional neural network; receptive fields; spatial transformer

Introduction

Neural encoding models aim to map sensory stimuli to neural responses and are a central tool in systems neuroscience (van Gerven, 2017). A popular strategy is to extract features from deep networks trained on vision tasks and map them to neural activity using a linear readout (Yamins et al., 2014; Güçlü & van Gerven, 2015; St-Yves & Naselaris, 2018; Khosla, Jamison, Kuceyeski, & Sabuncu, 2022; Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018; Kay, Naselaris, Prenger, & Gallant, 2008). However, many of these models ignore key spatial constraints observed in biological vision.

Recent work incorporates biologically motivated inductive biases such as spatial locality and structured receptive fields (Kietzmann et al., 2019; Banino et al., 2018; Khosla, Ngo, Jamison, Kuceyeski, & Sabuncu, 2020). For example, fwRF models (St-Yves & Naselaris, 2018) and spatial-feature decoupling (Wang et al., 2020) encourage more interpretable feature mappings.

We introduce the **Affine Feature Response Transform (AFRT)**, a compact model that learns a small spatial crop as a stand-in for a neuron's receptive field. This crop is used to extract localized CNN features. Each transformation is defined by simple shift and scale parameters, which makes the model both efficient and easy to interpret.

Methods

We model each neural response using features from a localized region of the image, rather than the full frame. A learnable affine transformation—parameterized by horizontal shift (t_x), vertical shift (t_y), and isotropic scale (s)—selects a region, which is then passed through a pretrained CNN feature extractor. This mimics the spatial specificity of receptive fields in visual cortex.

The cropped region is resampled via bilinear interpolation and processed by the CNN. Features are reduced to a $1 \times 1 \times C$ vector and linearly mapped to predict the neural response. Formally, for image s and transformation T, the predicted response is $f_\theta(s) = w^\top \phi(T(s))$, where ϕ denotes the CNN and w are linear weights. Both w and T are optimized via gradient descent to minimize mean squared error.

This approach reduces per-neuron parameters to under 7,000, improving efficiency and interpretability by constraining the model to spatially meaningful input regions.

Experimental data We used the THINGS ventral stream spiking dataset (TVSD) (Papale, Wang, Self, & Roelfsema, 2025), comprising recordings from 1024 electrodes (V1, V4, IT) in a macaque viewing 25,248 naturalistic images from the THINGS database (Hebart et al., 2019). Images (500×500 px) were shown in rapid 4-image sequences (200 ms on/off), shifted to the lower-right fovea.

Multi-unit activity (MUA) was time-averaged and normalized per channel (Burns, Xing, & Shapley, 2010). Reliability was assessed via pairwise correlations across 30 repetitions of 100 test images, and channels with mean reliability < 0.4 were excluded, resulting in 667 signals used for decoding.

Modeling and Evaluation We trained our Affine Feature Response Transform (AFRT) model on the TVSD dataset. AFRT applies a learnable spatial transformation to each stimulus, followed by feature extraction using a pretrained AlexNet (Güçlü & van Gerven, 2015). Features from Conv1–5 are linearly mapped to neural responses. A linear baseline model without spatial warping was also trained for comparison.

Models were trained on 22,348 stimuli and evaluated on 100 repeated test images using Pearson correlation. For each MUA signal, the best-performing layer (Conv1, Conv2, or Conv5) was selected, and the corresponding affine-transformed region defined the effective receptive field.

Results and Discussion

We evaluated AFRT on multi-unit activity (MUA) recorded from macaque visual areas V1, V4, and IT during natural image viewing. Compared to conventional CNN-based encoding models, AFRT improved prediction accuracy while significantly reducing the number of learnable parameters per neuron.

AFRT improves accuracy with fewer parameters

Despite relying on compact, site-specific spatial transformations, AFRT outperforms strong baselines in predicting MUA responses. Figure 2 compares test-set correlation scores between AFRT and two baselines across all brain areas. Each point corresponds to a single MUA site.

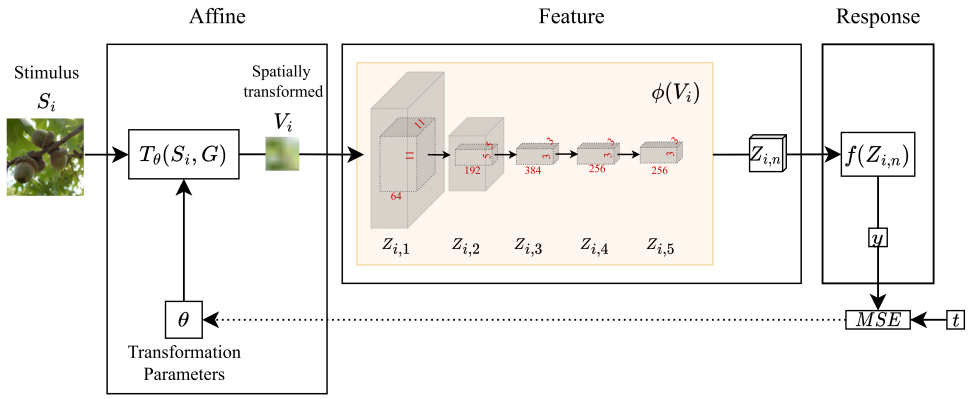

Figure 1: Schematic overview of the AFRT model. Each neuron learns a spatial crop of the input image via a learnable affine transform (scale and translations). The transformed image is passed through a CNN and spatially collapsed, and the resulting feature vector is linearly mapped to predict the response.

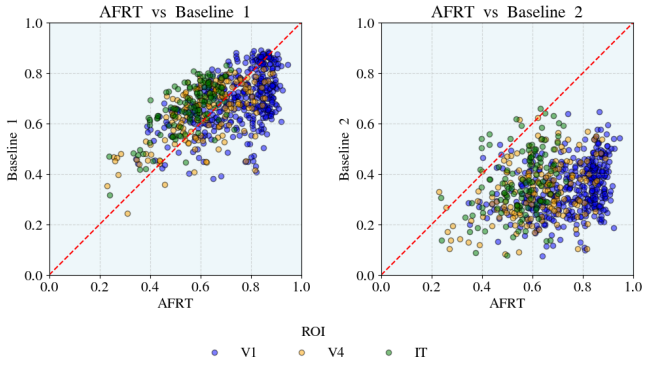

Figure 2: MUA site-wise test-set correlation for AFRT vs. baselines. Left: AFRT vs PCA-based baseline. Right: AFRT vs bilinear-resized input baseline. Each point is one MUA site across V1, V4, and IT. Red dashed line indicates equal performance.

The left panel compares AFRT to a model that uses PCA-reduced AlexNet features and a linear response layer. The right panel compares AFRT to a baseline using bilinear-resized inputs. In both comparisons, AFRT consistently achieves equal or better performance, particularly for low-SNR sites. While baseline models use tens or hundreds of thousands of parameters per site, AFRT achieves these results using fewer than 7,000 parameters by enforcing spatial locality.

AFRT recovers hierarchical receptive field structure

To assess whether AFRT captures biologically plausible spatial structure, we visualized the learned affine transformations as receptive field crops. These image regions correspond to the most predictive spatial input per site. As shown in Figure 3, receptive fields become larger and more diffuse from

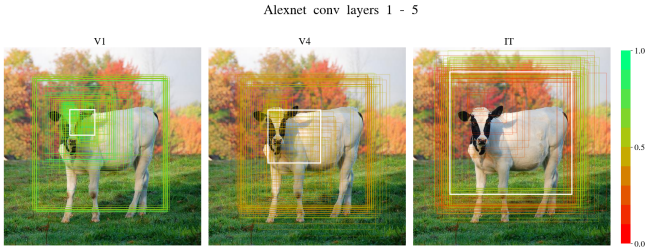

Figure 3: Learned receptive field crops across brain regions. Each square is one site's affine crop; color encodes model performance (Pearson R on the test set). White squares show regional means.

V1 to IT, mirroring known receptive field scaling in the ventral stream.

Notably, this retinotopic organization emerges without any spatial supervision. Early visual areas (e.g., V1) yield small, tightly localized crops, while downstream regions (e.g., IT) exhibit broader, overlapping selections. This pattern validates AFRT's spatial inductive bias.

Occasionally, large receptive fields appear in V1 as well, possibly reflecting the fact that MUA signals pool over small neural populations. In such cases, AFRT may learn to attend to a broader region to capture shared or overlapping activity.

Summary

AFRT improves neural encoding performance while dramatically reducing model complexity. Its inductive bias toward localized spatial input supports both predictive accuracy and biological plausibility. Despite being simple and lightweight, the model recovers known organizational principles of the visual cortex, making it a compelling approach for large-scale neural system identification tasks.

References

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., ... et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433.

Burns, S. P., Xing, D., & Shapley, R. M. (2010). Comparisons of the dynamics of local field potential and multiunit activity signals in macaque visual cortex. *Journal of Neuroscience*, *30*(41), 13739–13749.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS ONE*, *14*(10), e0223792.

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

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.

Khosla, M., Jamison, K., Kuceyeski, A., & Sabuncu, M. (2022). Characterizing the ventral visual stream with response-optimized neural encoding models. *Advances in Neural Information Processing Systems*, *35*, 9389–9402.

Khosla, M., Ngo, G., Jamison, K., Kuceyeski, A., & Sabuncu, M. (2020). Neural encoding with visual attention. *Advances in Neural Information Processing Systems*, *33*, 15942–15953.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Papale, P., Wang, F., Self, M. W., & Roelfsema, P. R. (2025). An extensive dataset of spiking activity to reveal the syntax of the ventral stream. *Neuron*.

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202.

van Gerven, M. A. J. (2017). A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*, *76*, 172–183.

Wang, H., Huang, L., Du, C., Li, D., Wang, B., & He, H. (2020). Neural encoding for human visual cortex with deep neural networks learning "what" and "where". *IEEE Transactions on Cognitive and Developmental Systems*, *13*(4), 827–840.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.